UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

EDUARDO GABRIEL CORTES

Beyond Accuracy: Completeness and Relevance Metrics for Evaluating Long Answers

Thesis presented in partial fulfillment of the requirements for the degree of Doctor of Computer Science

Advisor: Prof. Dr. Dante Augusto Couto Barone
Co-advisor: Prof. Dra. Renata Vieira

Porto Alegre
maio 2024

CIP — CATALOGING-IN-PUBLICATION

Cortes, Eduardo Gabriel

Beyond Accuracy: Completeness and Relevance Metrics for Evaluating Long Answers / Eduardo Gabriel Cortes. – Porto Alegre: PPGC da UFRGS, 2024.

165 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2024. Advisor: Dante Augusto Couto Barone; Co-advisor: Renata Vieira.

1. Question answering. 2. Non-factoid questions. 3. Long answers. 4. Answer evaluation. 5. Systematic review. I. Barone, Dante Augusto Couto. II. Vieira, Renata. III. Título.

ABSTRACT

The development of Question Answering (QA) systems that provide long answers face significant challenges in assessing the quality of these answers. Developing metrics capable of evaluating specific criteria individually, such as completeness, relevance, correctness and comprehensiveness, are important for identifying weaknesses and guiding improvements in these systems. Traditional metrics, like BLEU and ROUGE, often fail to capture semantic details and linguistic flexibility, and rely on a single score value that indicates how similar the system generated answer is compared to a reference answer. In this context, the goal of this work is to initiate and establish research, development, and validation of specific metrics to evaluate the completeness and relevance of answers provided by QA systems. For this purpose, a systematic review of non-factoid QA systems was conducted, followed by the creation of a dataset specifically annotated to assess completeness and relevance, containing long answers annotated by humans based on these criteria. Three metric models for evaluating these criteria were proposed: a prompt-based strategy using Large Language Models (LLMs), such as GPT-4; a model that adapts concepts of precision and recall to assess relevance and completeness, respectively, by segmenting the answer into discrete information units; and a regression model trained with synthetic data to assign scores of completeness and relevance. The experiments conducted compared these new metrics with conventional metrics to assess their correlation with human evaluations. The results highlighted the efficacy of the prompt model with GPT-4, which showed high correlation with human judgment, as well as the regression model, which shows high correlation in evaluating completeness, suggesting that metrics that do not require reference answers are competitive and can surpass traditional metrics in various scenarios.

Keywords: Question answering. Non-factoid questions. Long answers. Answer evaluation. Systematic review.

Além da Acurácia: Métricas de Completude e Relevância para Avaliar Respostas Longas

RESUMO

O desenvolvimento de sistemas de *Question Answering* (QA) que fornecem respostas longas enfrenta desafios significativos na avaliação da qualidade dessas respostas. Desenvolver métricas capazes de avaliar critérios específicos individualmente, como completude, relevância, correção e abrangência, é importante para identificar fraquezas e orientar melhorias nesses sistemas. Métricas tradicionais, como BLEU e ROUGE, muitas vezes falham em capturar detalhes semânticos e flexibilidade linguística, e dependem de um único valor de pontuação que indica o quanto a resposta gerada pelo sistema é semelhante a uma resposta de referência. Neste contexto, o objetivo deste trabalho é iniciar e estabelecer pesquisa, desenvolvimento e validação de métricas específicas para avaliar a completude e relevância das respostas fornecidas por sistemas de QA. Para esse fim, foi realizada uma revisão sistemática de sistemas de QA não-factoides, seguida pela criação de um conjunto de dados especificamente anotado para avaliar completude e relevância, contendo respostas longas anotadas por humanos baseadas nestes critérios. Foram propostos três modelos de métricas para avaliar esses critérios: uma estratégia baseada em prompts usando Large Language Models (LLMs), como o GPT-4; um modelo que adapta conceitos de precisão e revocação para avaliar relevância e completude, respectivamente, segmentando a resposta em unidades discretas de informação; e um modelo de regressão treinado com dados sintéticos para atribuir pontuações de completude e relevância. Os experimentos realizados compararam essas novas métricas com métricas convencionais para avaliar sua correlação com avaliações humanas. Os resultados destacaram a eficácia do modelo de prompt com GPT-4, que mostrou alta correlação com o julgamento humano, bem como o modelo de regressão, que mostra alta correlação na avaliação de completude, sugerindo que métricas que não requerem respostas de referência são competitivas e podem superar métricas tradicionais em vários cenários.

Palavras-chave: *Question answering*. Perguntas não fáticas. Respostas longas. Avaliação de resposta. Revisão sistemática.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ICC	Intraclass Correlation Coefficient
KG	Knowledge Graph
LLM	Large Language Model
NER	Named Entity Recognition
NLG	Natural Language Generation
QA	Question Answering

CONTENTS

1 INTRODUCTION

Question Answering (QA) systems are designed to answer questions in natural language, providing precise and informative answers. Questions that require lengthy answers, such as “Why is the sky blue?”, introduce additional complexity to these systems. Furthermore, evaluating the long answers from these systems is a challenging task because the longer the text to be evaluated, the greater the possibilities for expressing this information, which complicates its comparison with a reference answer, which may be semantically identical but structurally distinct.

Automatic text evaluation metrics, such as BLEU and BERTScore, are used to quantify the similarity between the provided answer and a reference answer through a single numerical value. However, these metrics aims to synthesize multiple characteristics, such as accuracy, completeness, relevance, and fluency, into a single measure, which can make it difficult to adequately evaluate each distinct characteristic.

This work proposes an analysis specifically focused on the criteria of completeness and relevance in long answers generated by QA systems. It seeks to develop a more refined methodology to evaluate these criteria, through the application of automatic metrics, allowing for a deeper and more detailed understanding of the effectiveness of QA systems.

1.1 Motivation

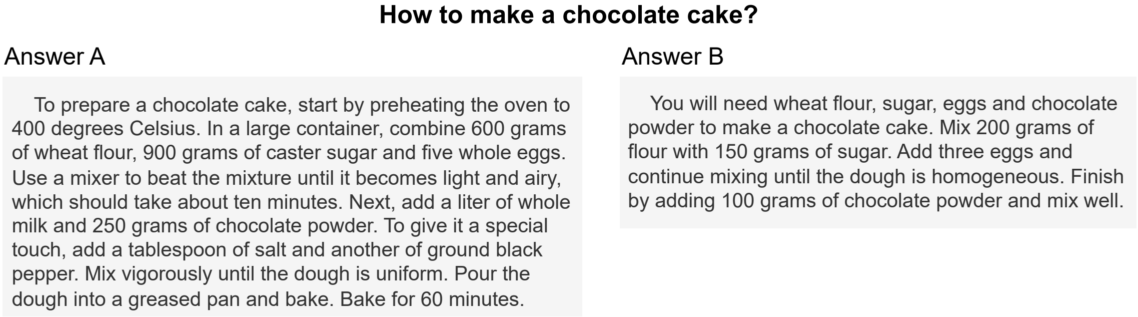
QA systems that provide long answers typically contain mechanisms capable of generating natural language (DENG et al., 2023) or mechanisms that provide an answers extracted directly text segments in documents (DARVISHI et al., 2023). To validate their performance, various questions are submitted to the system and their answers are checked, often comparing them with a reference answer. This evaluation process can be done manually, where human evaluators check the system’s answers, assigning scores for each answer (KHILJI et al., 2021). However, manual evaluation faces challenges such as the high demand for human effort for assessment, especially with long answers, where the effort is greater due to the larger amount of text to be analyzed.

Automatic methods for evaluating long answers involve using automatic metrics, which typically determine the similarity of the answer generated by a QA system in relation to a reference answer, considered correct. This type of evaluation has the advantage of not requiring human efforts for assessment. However, traditional automatic metrics such as BLEU (PAPINENI et al., 2002) and ROUGE (LIN, 2004) present significant difficulties, particularly in terms of dealing with semantics and linguistic flexibility (DEUTSCH; ROTH, 2021; ISABELLE; CHERRY; FOSTER, 2017; SULEM; ABEND; RAPPOPORT, 2018). As these metrics are based on n-gram overlap, they primarily measure the literal similarity of texts, ignoring whether two sentences have the same meaning but are phrased differently.

More recent metrics, such as BERTScore (ZHANG* et al., 2020) and BARTScore (YUAN; NEUBIG; LIU, 2021), provide a more accurate assessment of semantic similarity between texts, addressing many of the limitations found in traditional metrics such as BLEU and ROUGE. These newer metrics use transformer language models, which are capable of understanding broader contexts and semantic details. However, these metric models still face challenges, such as the requirement for greater computational power due to the use of large and complex language models and struggle with adversarial attacks (deliberate inputs to mislead models) related to lexical overlap and inaccuracies in content (CHEN; EGER, 2023). Additionally, understanding how these scores are derived from the models can be challenging and requires a deeper knowledge of how language models work.

A significant limitation of similarity metrics for evaluating long answers is the need to have a reference answer for each entry to be evaluated. This can be problematic, as it requires the availability of a set of “golden answer”, which must be carefully reviewed and validated by experts to ensure their quality, which not only increases the cost and time required to develop these sets, but also introduces a characteristic of subjectivity in the evaluation process (KEELER, 2011). Therefore, while these metrics offer an automated form of evaluation, the dependence on reference answers may limit their applicability in different contexts, as in cases where there is a dataset with quality reference answer.

Using a single score to determine the quality of an answer allows for direct performance comparison between QA systems. However, this overall quality score does not clearly present the performance of these models on specific quality criteria of the generated answers. For instance, an answer may be complete and clear but contain inaccurate information, as shown in example (A) of Figure 1.1. Conversely, an answer may contain accurate and relevant information but be incomplete, as shown in example (B) of Figure 1.1. Therefore, instead of using a single numerical value to represent various quality criteria, metrics that evaluate each criterion individually may allow for a more detailed and



How to make a chocolate cake?

Answer A

To prepare a chocolate cake, start by preheating the oven to 400 degrees Celsius. In a large container, combine 600 grams of wheat flour, 900 grams of caster sugar and five whole eggs. Use a mixer to beat the mixture until it becomes light and airy, which should take about ten minutes. Next, add a liter of whole milk and 250 grams of chocolate powder. To give it a special touch, add a tablespoon of salt and another of ground black pepper. Mix vigorously until the dough is uniform. Pour the dough into a greased pan and bake. Bake for 60 minutes.

Answer B

You will need wheat flour, sugar, eggs and chocolate powder to make a chocolate cake. Mix 200 grams of flour with 150 grams of sugar. Add three eggs and continue mixing until the dough is homogeneous. Finish by adding 100 grams of chocolate powder and mix well.

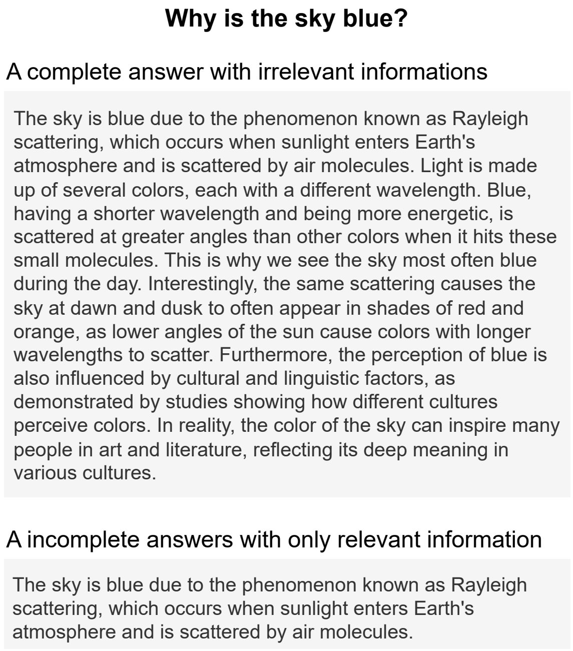
Figure 1.1 – Example of two answers to the same question, where answer A is complete and clear, but contains inaccurate information, such as high degrees Celsius for the roast, a high amount of sugar, and inappropriate ingredients, such as pepper. On the other hand, answer B is accurate and incomplete, lacking instructions on what to do after mixing the ingredients.
Source: The Author.

specialized analysis, offering a more granular view of the quality of the answers. This can help QA system developers to identify specific areas for improvement and refine their approaches as needed.

There are different individual criteria that can be evaluated in a long answer. According to the study by (CAMBAZOGLU et al., 2021), which assesses the aspects that make an answer useful, the criteria of correctness, relevance, and completeness are the most important aspects and have a relatively high correlation with the perceived usefulness of the answers given:

- **Correctness:** also known as accuracy, represents how accurate the content of the answer is, i.e., whether the answer contains correct and true information. An answer is considered accurate when it provides a true and correct answer to the question;
- **Relevance:** concerns the suitability of the answer to the context of the question. An answer is relevant if it directly addresses the topic that was asked, being directly related to the subject of interest to the user. This means that the answer should focus on what was requested in the question, without deviating to unsolicited related information;
- **Completeness:** is related to how complete the answer is in covering all essential aspects of the question. A complete answer should not leave important aspects of the question not answered, offering a complete answer that fully satisfies the user's information need.

With the development of generative LLMs, such as GPT-4, the ability of systems to provide answers to instructions has improved significantly (BROWN et al., 2020; OPENAI et al., 2024; LEWIS et al., 2019). Given this, metrics capable of assessing the



Why is the sky blue?

A complete answer with irrelevant informations

The sky is blue due to the phenomenon known as Rayleigh scattering, which occurs when sunlight enters Earth's atmosphere and is scattered by air molecules. Light is made up of several colors, each with a different wavelength. Blue, having a shorter wavelength and being more energetic, is scattered at greater angles than other colors when it hits these small molecules. This is why we see the sky most often blue during the day. Interestingly, the same scattering causes the sky at dawn and dusk to often appear in shades of red and orange, as lower angles of the sun cause colors with longer wavelengths to scatter. Furthermore, the perception of blue is also influenced by cultural and linguistic factors, as demonstrated by studies showing how different cultures perceive colors. In reality, the color of the sky can inspire many people in art and literature, reflecting its deep meaning in various cultures.

A incomplete answers with only relevant information

The sky is blue due to the phenomenon known as Rayleigh scattering, which occurs when sunlight enters Earth's atmosphere and is scattered by air molecules.

Figure 1.2 – Examples of a complete answer, but with irrelevant information, and another relevance answer, however, lacking relevant information.
Source: The Author.

completeness and relevance of the generated answers offer the ability to verify whether these models are generating information that covers all requested topics and also aligns with the informational needs of the users.

There is a relationship between completeness and relevance, which can be observed in both long and short answers, as seen in Figure 1.2. Long answers tend to cover more aspects of a question, possibly achieving a high completeness score. On the other hand, shorter answers must have a higher relevance, focusing strictly on what was requested, without adding unnecessary information. Thus, if a system aims to generate longer and more complete answers, there is an increased possibility of these answers containing information irrelevant to the user. In the other hand, if the system aims to provide shorter and more relevant answers, there is an increased chance these answers will be incomplete for the user. Therefore, QA systems can be designed with a specific focus: some prioritize generating complete and detailed answers, while others concentrate on maximizing the relevance of the content provided, each approach with its respective advantages and disadvantages.

Accuracy, or correctness, although critical, requires depth analysis in fact-checking (MIN et al., 2023; AHARONI et al., 2023; HONOVICH et al., 2022), extending beyond the scope of this study. Instead, this research focuses on the completeness and relevance criterias. Also, it is important to recognize that the perception of completeness and relevance can be subjective, varying according to the expectations and information needs of the user. For example, a person *A* with knowledge of NLP requires less information to understand "How does the BLEU metric work?" than a person *B* who is new to the subject. In other words, for person *A*, an answer with many explanations tends to have more irrelevant information. While an answer with little explanation tends to be more incomplete for person *B*.

The importance of this research extends beyond the technical issues of QA systems, potentially reaching social, business, and academic implications. Socially, by enhancing the ability of QA systems to provide complete and relevant answers, we can democratize access to quality information, benefiting especially educational contexts with limited access to informational resources. In the business context, more efficient QA systems help in various sectors, such as customer service, providing answers that satisfy user needs. From an academic perspective, this investigation directly contributes to the evolution of evaluation metrics, encouraging the development of more refined and specific approaches that brings advancements in the field of NLP.

1.2 Objectives

The general objective of this research is to initiate the development and validation of metrics specifically designed to assess the completeness and relevance of long answers provided by QA systems. This objective is supported by the lack of existing metrics for such evaluation. Also, it seeks to establish the foundation for developing new differentiated metrics to evaluate these specific criterias that align with human judgment. The specific objectives of this thesis are described as follows:

- **Systematic Analysis of Non-Factoid QA Systems:** Conduct a comprehensive review of non-factoid QA systems to discern the various tasks, methods, data sets, evaluation strategies, and outcomes pertinent to generating long answers. This analysis highlighted the key areas for improvement and underline the need for improvement in metrics that can evaluate long answers.
- **Development of an Annotated Database:** Create an initial and annotated database that facilitates the empirical evaluation of metric models concerning completeness and relevance criteria. This database will serve as a resource for testing and refining evaluation metrics, aiming to be reliable in measuring the qualities that define

completeness and relevence in answers.

- **Proposal of Metric Models for Completeness and Relevance:** Develop novel metric models that focus on the assessment of long answers for completeness and relevance criteria, emphasizing:
    - A prompt-based strategy that utilizes a generative LLM to evaluate of long answers, leveraging the model's capability to interpret and analyze text comprehensively.
    - A strategy that adapts the concepts of recall and precision to assess completeness and relevance, respectively, by segmenting the evaluated answer into discrete information units.
    - A regression model trained on a synthetic dataset to assign completeness and relevance scores.
- **Evaluation of Conventional Metrics:** Assess how the current conventional metrics approximate the evaluation of completeness and relevance in long answers. This objective aims to benchmark these traditional metrics against the newly proposed models to establish a baseline for improvement.

1.3 Contributions

This thesis brings advancements in the field of QA for long answers, proposing new approaches and evaluation techniques that specifically focus on the criteria of completeness and relevance. The main contributions of this work are listed below:

- **Systematic Review on Non-Factoid QA**: The review conducted offers a comprehensive analysis of non-factoid QA systems, presenting methods, tasks, datasets, evaluation strategies, and outcomes obtained. This analysis highlights the complexity of longer answers and the necessity for evaluation methods. The review identifies gaps in the existing literature, especially in the systems' ability to compose detailed answers and consider context from various information sources. Thus, this systematic review not only summarizes the state of the art in non-factoid QA but also establishes a starting point for future developments in the field, as seen in this work.
- **Dataset for Metric Evaluation:** One of the main contributions of this work is the

creation of an annotated dataset focused on “Instruction” type questions in the field of Computer Science, where the answers are lengthy and were annotated by humans based on completeness and relevance criteria. This dataset can be used as a tool for evaluating metric models that focus on these criteria, aiming for a controlled environment to minimize biases and help the interpretability of the results. This annotated database can be used as a resource for testing and refining evaluation metrics. Also, it can be used as a foundation for future research aimed at better understanding how long answers relate to the completeness and relevance criterias.

- **Proposed Metric Models to Evaluate Completeness and Relevance:** This work contributes to the development of novel metric models specifically designed to assess the completeness and relevance of long answers. The metrics aim to understand how evaluated QA systems handle the depth and pertinence of the information they provide, addressing two critical dimensions that directly influence the usefulness of the answers to users. The proposed models include:

    - **Prompt-Based Strategy with Generative LLM:** Using the advanced text comprehension capabilities of LLMs, such as GPT-4, this model employs a prompt technique that guides the LLM to analyze and score the completeness and relevance of the answers.
    - **Adaptation of Recall and Precision Concepts:** This model adapts traditional metrics of recall and precision to measure, respectively, the completeness and relevance of the answers. By segmenting the answer into discrete information units, the model quantitatively assesses how much relevant information the answer contains in relation to what would be ideal (completeness) and how much of the answer’s content is relevant in relation to its total volume (relevance).
    - **Regression Model Based on Synthetic Data:** Developed to predict completeness and relevance scores, this model is trained with a synthetic dataset that simulates different levels of completeness and relevance. Using NLP techniques and models like BERT for text comprehension, the regression model is capable of assigning numerical values to the answers, quantifying their completeness and relevance.

    Each of these models offers a distinct approach to evaluating long answers, allowing an analysis of what have an acceptable performance for the continuous improvement of these evaluation criteria. Also, the proposed metric models designed to assess completeness and relevance do not require a reference answer to work. This independence from reference texts allows for a more flexible application in different contexts where such gold standards may not be available, thus, expanding the usefulness and applicability of the proposed metrics in the evaluation of long answers.

- **Evaluation and Comparison of Metrics for Completeness and Relevance Criteria:** This research contributes to the field of QA by evaluating and comparing different quality metrics for long answers, with a specific focus on completeness and relevance criteria. The analysis of both conventional metrics and the newly proposed metrics provides insights into how each aligns with human judgment, allowing an understanding of the capabilities and limitations of each evaluative method. Through the application of conventional metrics such as BLEU, ROUGE, BERTScore, and others, this research determines how much these metrics are capable of capturing the aspects of completeness and relevance of the answers.

With the proposal of new metric models that specifically focus on completeness and relevance, this research advances the state of the art by introducing methods that address the gaps left by traditional approaches. Each of the newly proposed models is tested and compared, not only among themselves but also against conventional metrics, to determine their effectiveness in replicating human evaluations.

1.4 Document Structure

In addition to the introduction, this thesis presents a theoretical description of QA field in Chapter 2, along with a discussion on the evaluation of long answers and related works that allow for the automatic evaluation of long answers. Chapter 3 presents the systematic review on non-factoid QA systems, showcasing the methods, tasks, datasets, evaluation strategies, and results obtained from these systems. In Chapter 4, the research methodology of this work is presented, detailing and justifying the choices of research stages. Chapter 5 describes the dataset constructed for the evaluation of metric models based on the criteria of completeness and relevance, along with analyses of the annotations made by humans. Subsequently, Chapter 6 introduces the three proposed metric models in this work focused on evaluating the completeness and relevance of long answers. Chapter

7 presents the results obtained and their analyses related to experiments with different metric models using the proposed dataset. Finally, Chapter 8 presents the conclusions, along with the limitations and future work.

2 BACKGROUND

This chapter provides an overview about the topic related with this research, landscaping mainly QA systems, emphasizing the architecture, classification, approaches, and evaluation metrics. It present the fundamental components of QA systems, explores different methods for classifying these systems, and highlights the importance of different question types. Furthermore, the chapter explore the aspect of evaluating QA systems, discussing metrics for assessing long answers. Through an analysis of related work, this chapter present how the automatic metrics work in evaluation long answers from QA systems, highlighting its challenges and the need for models that evaluate specific characteristics of answers, such as completeness and relevance.

2.1 Question Answering

QA systems is a field in Computer Science, aimed at automatically providing precise answers to question in natural language. To correctly answer, these systems face challenges that involve NLP, Machine Learning, Information Retrieval, and Information Extraction (SASIKUMAR; SINDHU, 2014). QA include various NLP tasks associated with understanding and generating natural language. This section present the structure and components of QA systems, highlighting the processes of understanding, retrieving, and generating natural language answers.

One of the first QA systems was the *BASEBALL* (JR et al., 1961), a restricted domain system which answered questions about a single season of the sport. Also, the 1977 *LUNAR* system allowed geologists to ask questions about the domain of rocks. More recently, IBM's widely-dominated Watson system managed the feat of beating humans at the *game-show Jeopardy!* in 2011. Actually, models are outperforming the human performance in some benchmarks, as the system (ZHANG; YANG; ZHAO, 2020) in the SQuAD 2.0 benchmark (RAJPURKAR; JIA; LIANG, 2018) and (XIA; WU; YAN, 2019) in the MS-MARCO benchmark (BAJAJ et al., 2018).

2.1.1 Classifying QA systems

QA systems are remarkably diverse, categorized based on the complexity of natural language, knowledge domains, and the forms of information resources they utilize (SOARES; PARREIRAS, 2020; DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019). These differences are relate to the complexity of natural language, the large possibilities of knowledge domains, and the depth of detail presented, as well as the different forms of information resources. Therefore, developing a system that masters the entire complexity of natural language, capable of handling all types of data structures, and with deep knowledge in all areas, remains a challenge. Consequently, there are different types of QA systems built for specific aspects. This section aims to organize the various categories of QA systems through three forms of categorization based on the system's input question type, the type of knowledge source, and the knowledge domain of the systems.

2.1.1.1 Question Type in QA Systems

Natural language questions are a discourse form aimed at information retrieval. Each question type serves a specific purpose and requires a distinct type of answer. Questions can demand a concise factoid piece of information, such as "What is the capital of Brazil?" which only needs the name of a city as an answer. Others may require a more elaborate explanation, such as "Why is the sky blue?", which demands a descriptive answer. The question type can vary not just in answer length but also in the reasoning steps required to provide an answer. For example, the question "How many goals did the 2020 Champions League winning team score?" first requires identifying the 2020 Champions League winner and then calculating their total goals scored during the tournament.

Depending on the question type, the functionality and complexity of the QA system can significantly change. For instance, questions seeking short factoid information might simply extract the requested information from a text document, whereas a question asking for a comparison of two literary works necessitates analyzing two distinct books to generate an answer. Thus, the type of question a system aims to answer plays a significant role in its development.

Different studies propose taxonomies to categorize questions into various types. One of the main frameworks is presented by (LI; ROTH, 2002), presenting a two-level granularity taxonomy with several categories at each level. The first level provides a more abstract categorization, such as "ENTITY", while the second level details specific subcategories, like “Animal”. For instance, a question like “Who is the current president of Portugal?” would be classified under “HUMAN” and “Individual” as the expected answer is a person’s name, whereas “What is a prism?” would fall under “DESCRIPTION” and “Definition” expecting a text defining what a prism is.

A broader division of questions can be into factoid and non-factoid groups (YANG et al., 2019). Factoid questions have the characteristic to require a fact as the answer, such as a name, a location, or a date (YANG et al., 2019). For example, the question “What is the capital of Italy?” requires a location as the answer. On the other hand, non-factoids represent questions that do not require a simple fact as answers. Usually, these questions require more extensive and complex information as the answer. For example, the question “What are the advantages of purchasing auto insurance?” expect the answer as a long text compared with a factoid question.

- **Factoid**: These questions require returning a single fact as the answer. Examples include “When did World War II start?” or “What is the capital of Spain?”. Most categories in (LI; ROTH, 2002) taxonomy can be considered types of factoid questions, except for the “DESCRIPTION” category.
- **Non-factoid**: These typically demand longer answers, such as descriptions, opinions, or explanations. For instance, “Why is the sky blue?” or “What are the steps to earning a master’s degree?”.

Table 2.1 outlines different types of non-factoid questions, highlighting distinct challenges for the QA system. Beyond these categories, there can be various other question types within the factoid and non-factoid groups, including those requiring lists (“What are the world’s largest cities?”), hypothetical scenarios (“What would happen if the moon disappeared?”), or based in example answers (“How does art influence society?”).

2.1.1.2 Knowledge Source Types in QA Systems

A fundamental component of a QA system is the knowledge source used to extract the necessary information to answer the input question. Knowledge sources can vary in how knowledge is created, stored, modified, and queried.

Unstructured data, that does not have a standard structure, is easily created but presents challenges in processing and analyzing due to the requirement for advanced natural language comprehension. Examples include raw text documents and web pages,

Type	**Description**	**Example**
Definition	Questions requiring the definition of something. Usually start with "What is ..."	"What is a prism?"
Explanatory	Questions requiring an explanation or context. Usually start with "Why ..."	"Why is the sky blue?"
Procedural	Questions requiring a set of steps or instructions to do something. Usually start with "How ..."	"How to make a chocolate cake?"
Comparative	Questions requiring a comparison between two or more subjects.	"What are the differences between SSD and HDD?"
Opinion	Questions requiring a personal perspective or evaluation.	"What do you think about modern art?"
Confirmation	Questions requiring a "Yes" or "No" answer.	"Is Athens the capital of Greece?"

Table 2.1 – Examples of non-factoid questions with examples.
Source: The Author.

which, despite HTML tags, don't provide a deep enough structure for detailed information representation. This data type necessitates robust retrieval and NLP techniques to extract relevant information.

In contrast, structured data sources follow a standard format that simplifies data access and interpretation by the system. This category includes relational databases and knowledge graphs, which organize information in a way that facilitates semantic understanding essential for QA. Structured data allows for direct queries (e.g., SQL for databases) to efficiently retrieve specific information, enhancing the system's ability to provide accurate answers.

Currently, there is parametric memory which can also be considered a form of knowledge representation of a QA system. It represents information through embedding knowledge directly within the architecture of deep learning models. These models, especially those based on transformers like BERT (DEVLIN et al., 2019) and GPT (BROWN et al., 2020; ACHIAM et al., 2023), uses a large number of trainable parameters to store and apply learned information to answer questions. However, the "black box" nature of these models have challenges in interpretability and reliability, as the knowledge is not directly accessible or understandable in human terms. Combining end-to-end transformer models with other approaches, such as providing relevant context or integrating reliable information sources, can improve the accuracy and reliability of the answers. This hybrid strategy addresses challenges such as model hallucination, where answer information is based on true and verified content.

2.1.1.3 Knowledge Domain in QA Systems

QA systems can be categorized based on the knowledge domain of knowledge in which the system will be applied. Some systems are designed to answer questions within specific fields, such as biomedicine or finance, while others aim to cover a broader range of topics. Therefore, these systems fall into two main groups: open domain and closed domain.

Open domain systems are designed to answer questions among a wide range of topics. Due to the extensive variety of knowledge required, these systems often use unstructured knowledge sources such as raw text documents and web search engines, which provide a vast amount of data. However, one of the most important challenge with open domain systems is to ensure the precision and relevance of the answers due to the vast possibility of information, necessitating sophisticated data retrieval and interpretation processes.

Closed domain systems focus on a specific topic, allowing for a more detailed exploration and understanding of the the subject. Structured data sources are commonly used with closed domain systems to provide detailed and specific information, once this kind of knowledge source can delivere more accurate information. While closed domain systems benefit from a deeper and more specialized understanding, they face challenges in staying updated with information from their topic, requiring continuous data structuring and updating efforts.

2.1.2 QA approaches

This section describes the approaches to QA, detailing the typical architecture of document-based QA systems and those based on Knowledge Graphs or end-to-end models. It illustrates the sequential stages of Question Processing, Information Retrieval, and Answer Processing, presenting the complexities involved in each phase.

As observed in QA studies (KODRA; KAJO, 2017; SOARES; PARREIRAS, 2020; DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019; NORASET; LOWPHANSIRIKUL; TUAROB, 2021), factoid and non-factoid QA systems present a similar architecture, which is based on three key components: 1) Question Processing, 2) Informa-

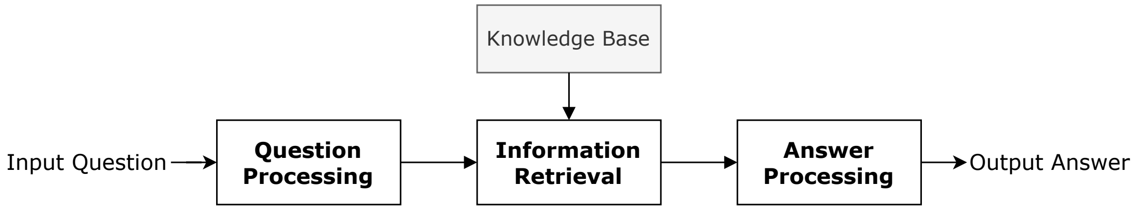


Figure 2.1 – Question Answering System Architecture.
Source: The Author.

tion Retrieval, and 3) Answer Processing. While Question Processing and Information Retrieval have similar behavior for both factoid and non-factoid QA systems, Answer Processing presents the most significant difference. Figure 2.1 shows the general architecture of QA systems. However, the studies do not always design their systems following precisely this architecture. The architecture here presented shows the main components and tasks of general architecture, mainly for document-based QA systems, following previously works (KODRA; KAJO, 2017; SOARES; PARREIRAS, 2020; DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019; YOGISH; MANJUNATH; HEGADI, 2018; LIU et al., 2016; CHALI; HASAN; MOJAHID, 2015). Regarding the terminology used to describe the components and tasks, it tries to simplify and cover as many problems as possible since QA systems' development involves several challenges.

2.1.2.1 Document-based QA Systems

QA systems that operate based on textual documents as the primary source of information are designed to answer questions asked by users by retrieving and processing information from a predefined set of documents (LI; LI; WU, 2018; ZHU et al., 2021). The architecture of these systems can be understood as a modular process, where each module is responsible for a specific step to answer the input question. These steps include Question Processing, Information Retrieval and Answer Processing. Figure 2.1 illustrates the typical architecture of a document-based QA system.

The **Question Processing** stage is an important initial step in QA systems, with the main purpose is to interpret the user's intent and to facilitate the next processing phases. This stage may include several distinct tasks, such as keyword extraction, question classification, answer reformulation (HERMJAKOB; ECHIHABI; MARCU, 2002), among others possible (CORTES et al., 2020; CORTES et al., 2022). For instance, keyword extraction focuses on identifying key terms in the question that are important for the search for information. This task, using methods like tokenization and stop-word elimination, seeks to refine the query for information retrieval.

In parallel, another example of a task is the classification of the question, which involves categorizing the question into a specific type, such as a question that requires a “Person” as an answer (“Who discovered Brazil?”) or a “Date” (“When was Pedro Álvares Cabral born?”). There is also the possibility of classification for non-factoid questions, which require detailed or explanatory answers (non-factoid), such as “How does photosynthesis work?” or “Explain the causes of World War I”, this classification can include categories like “Process” or “Explanation”. Such classifications facilitate the selection of information, aims that the answers provided meet the user’s expectations (WU et al., 2015; Ben Abacha; ZWEIGENBAUM, 2015; BONDARENKO et al., 2020; CORTES; WOLOSZYN; BARONE, 2018).

The classification technique usally uses supervised machine learning models, recognized for their ability to determine the class based on the text of the question. These models are trained by large annotated datasets, which have a wide range of question types and their corresponding classes (ABDEL-NABI; AWAJAN; ALI, 2023). In addition to supervised learning techniques, rule-based approaches can be also employed, particularly useful in scenarios with limitations of annotated data or when the questions present predictable structures (MADABUSHI; LEE, 2016).

The **Information Retrieval** stage serves as the mechanism by which relevant information is identified and extracted from the knowledge base. This process is typically hierarchical, progressively refining the search from entire documents to specific sentences containing the information needed to answer the user’s question. The steps involved in this process can involve tasks related to document retrieval, followed by the extraction of relevant paragraphs, and finally, the identification of key sentences. This process uses the terms and keywords identified in the question processing stage. Following the selection of documents, the focus narrows to the extraction of paragraphs and, subsequently, specific sentences that contain the desired information. This granular selection is important due to the possibility of the documents containing significant volumes of irrelevant information (SOARES; PARREIRAS, 2020).

The final stage in the QA process is the **Answer Processing**, which utilizes the information gathered and processed in the previous stages to produce the final answer to the user (DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019; YOGISH; MANJUNATH; HEGADI, 2018; PAPADAKIS; TZITZIKAS, 2015; LIU et al., 2016). This stage varies significantly depending on the type of question, adapting to provide both short (factual) and long (non-factoid) answers.

Initially, the extraction of candidate answers is conducted based on the data filtered during the Information Retrieval. This stage involves identifying elements in the text that helps to answer the question (AL-OMARI; DUWAIRI, 2023). For factual questions, this often means the identification and extraction of specific entities that match the question's category (for example, people, places, dates). This process can be done through Named Entity Recognition (NER) models and reading comprehension techniques, which are trained to extract the relevant information from the provided context (CORTES; WOLOSZYN; BARONE, 2018). For questions that demand more complex answers, approaches to summarization or text generation may be used, utilizing models that synthesize information into answers (LYU et al., 2021).

Following extraction, the ranking of candidate answers determines the most appropriate and accurate candidates. This process evaluates and scores each candidate answer based on criteria such as frequency, textual relevance, semantic similarity to the question, and, when applicable, the reliability of the source (LIN; WU; CHEN, 2021). The final result of this stage is the highest-scored answer, selected to be presented to the user.

2.1.2.2 QA Systems Based in Knowledge Graphs

There are also QA systems that employ Knowledge Graphs (KGs) as a source of information for retrieving and providing answers to questions (ZIRUI et al., 2021; HU et al., 2023; YASUNAGA et al., 2021; PEREIRA et al., 2022). KGs is structures that semantically encode knowledge through entities (nodes) and relationships (edges), allow for a rich and interconnected representation of information.

The implementation of these systems can be generalized into different key stages. For example, entity identification might be an initial stage that deals with classify the elements of the user's question that refer to known entities in the KG (JIANG; CHI; ZHAN, 2021). This is followed by entity linking, a stage in which each identified entity is associated with a corresponding node in the KG, a challenge that can be considerable due to variation in the representations of the same entity.

Query generation is a possible subsequent process, responsible for formulating a structured query, usually in SPARQL, from the original question, facilitating interaction with the KG (QIU et al., 2020). This phase may depend on the success of the previous stages, as precise entity linking is essential for constructing effective queries. Finally, in the answer generation phase, the system must translate the data retrieved from the KG into an answer understandable to the user (OMAR et al., 2023). This phase can vary considerably in complexity, from the simple enumeration of entities to the generation of descriptive and elaborate answers.

2.1.2.3 End-to-end QA Systems

There is also the possibility of an end-to-end QA system aimed at integrating all phases of the QA process, from question interpretation to answer generation, into a single model, typically consisting of a deep artificial neural network. This model is distinguished by its ability to process complex inputs, incorporating both the question and relevant context, often extracted from a reliable knowledge source (HE; GAO; CHEN, 2021; KIM et al., 2021; KIM; SON; KIM, 2021). The incorporation of context aims to mitigate limitations associated with the exclusive reliance on the model's parametric memory, which, although it can be considered efficient in certain aspects, can present challenges in accuracy and coherence.

The effective implementation of an end-to-end QA system is related to different phases, including data preprocessing, pre-training on extensive corpora to acquire a basic understanding of language, and specific training on datasets aligned with the QA task (CHAYBOUTI; SAGHE; SHABOU, 2021). These datasets are composed of question-answer pairs, along with relevant contexts that guide the model in generating appropriate answers.

The end-to-end model uses the capabilities of transformer architectures, which can be fine-tuned to perform both the generation of long answers and the handling of factual questions, reflecting the versatility of the approach. The use of pre-trained models specific to certain languages or domains allows for a more efficient initiation of the learning process, since these models already contain some prior knowledge of the language acquired in a previous training process (SACHAN et al., 2021).

Although the end-to-end approach represents the state of the art in QA, offering the ability to learn directly from question-answer examples, it also faces significant challenges (ABDEL-NABI; AWAJAN; ALI, 2023). These include the demand for large volumes of training data, the need for substantial computational power, and issues of interpretability and explainability of the generated results.

2.2 Evaluating QA System Answers

Evaluating the answers produced by QA systems is important for assessing their performance. This section explore the methodologies for evaluating long answers. It discusses conventional metrics like BLEU, ROUGE, and METEOR, and new ones, as BERTScore, an approach that uses deep neural networks for semantic analysis. The discussion extends to the challenges in evaluating long answers, showing the need for a multidimensional approach that assesses individual characteristics, as completeness and relevance.

Different methods for evaluating QA systems are better suited to the varied computational tasks involved in the processing stages of the system. For instance, classification stage evaluation can use methodologies applied in classification problems, while document retrieval stage evaluation is best approached with Information Retrieval evaluation methods. Direct evaluation of the system's output answers requires methods that vary according to the answer type, with some being more suitable for short, factoid answers, and others for longer answers.

For system comparison, evaluation metrics offer a quantitative basis for measuring performance by comparing the system's output with reference answers from the dataset (AMPLAYO et al., 2022; DEUTSCH; ROTH, 2022; CELIKYILMAZ; CLARK; GAO, 2021). These metrics differ according to the type of answer the QA system is designed to produce. For short and precise answers, such as a name, date, or specific fact, metrics typically assess the system's ability to correctly identify the exact answer within a set of candidate answers.

Long answers present a greater challenge for comparison and evaluation, especially for metrics based on word overlap (ISABELLE; CHERRY; FOSTER, 2017). The natural language allows for various ways to express information, so answers that express the same idea but with different words might score lower. For long answers, metrics determining textual similarity can be employed, including the traditional metrics, mainly used in translation tasks:

- **BLEU (Bilingual Evaluation Understudy):** Initially developed for machine translation evaluation, it compares the system's answer with one or more reference answers based on the overlap of n-grams (word sequences) between the generated answer and the reference answers (PAPINENI et al., 2002).
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Focuses more

on the system's ability to reproduce the content of reference answers. It uses n-gram overlap, longer subsequences, and word co-occurrences to evaluate automatic summaries, for instance (LIN, 2004).

- **METEOR (Metric for Evaluation of Translation with Explicit Ordering):** Similar to BLEU but more sophisticated as it also considers synonyms and the grammatical structure of the answers. METEOR aims to align more closely with human evaluation than BLEU (BANERJEE; LAVIE, 2005).

Metrics utilizing deep neural network models to predict the metric value aim to overcome this overlap issue through semantic analysis and contextualization of words within the answers. An example of this is BERTScore (ZHANG* et al., 2020), that aligns words between the system's and reference answers based on their embeddings, considering the context in which the words are used. This allows the metric to evaluate not just the exact word overlap but also semantic and contextual similarity. Thus, even if different words are used, if they share similar meanings in the given context, the answer can be evaluated positively. In addition to these, there is also the approach of using generative LLMs to evaluate answers, where a prompt is used with instructions on how to evaluate an answer (LI; PATEL; DU, 2023; KE et al., 2023; FAGGIOLI et al., 2023).

One of the major challenges in QA system evaluation is dealing with subjectivity, especially in long answers where "correctness" may be open to interpretation. Often, automatic metrics cannot handle such nuances. In these cases, qualitative metrics involving manual human evaluation of the system-provided answers can be employed. This may include:

- **Evaluation by Human Judges:** Human evaluators analyze the system-generated answers to judge specific criteria such as relevance, accuracy, and naturalness (MALAVIYA et al., 2024).
- **Usability Tests:** Observing how end users interact with the system and collecting their feedback on the effectiveness and usefulness of the answers (NAKANO et al., 2022).

These metrics are used to automatically evaluate long answers, focusing on the overall quality of the system's answer. The approach involves comparing the answer generated by the QA system with one or more reference answers. However, these metrics may not clearly show important individual aspects of the answer, such as the accuracy of the information provided, the fluency of the text, the completeness of the answer and its relevance to the question asked. For example, an answer can be highly fluent and grammatically correct but lack important information. Alternatively, an answer might contain all the necessary information but be presented in a way that's difficult to understand due to fluency or structure issues. The following subsection presents a discussion on metrics focused on evaluating individual characteristics of long answers.

2.2.1 Evaluating Specific Characteristics of Long Answers

Evaluation based solely on global comparisons may be insufficient to fully understand the strengths and weaknesses of a QA system. For example, a system might be very good at generating complete answers but may include many irrelevant details in the answer. Another system could be capable of generating answers that seem natural and are grammatically correct, but it might provide false information. Therefore, evaluating individual aspects of the answers provided by QA systems is important for understanding the overall effectiveness of these systems.

The ability of a QA system to provide answers that meet different characteristics may require a delicate balance. The overlapping competencies needed to generate answers that satisfy all these criteria are complex. For example, a system that seeks to provide complete answers will increase the size of the generated answers, but this characteristic increases the possibility that extra information will be irrelevant to the user, making the answers less relevant. On the other hand, if the system focuses on short and relevant answers, these may become incomplete. Therefore, it show that the evaluation of QA systems, especially for long answers, requires a more granular approach.

In this context, the studies by (FRICKÉ, 1997) and (BARRY; SCHAMBER, 1998) establish a theoretical framework for information evaluation. The study by (FRICKÉ, 1997) presents a vision for information evaluation that aligns with the concept of verisimilitude. It offers a theoretical framework that values the approximation of truth as fundamental criteria for information evaluation. The author suggests that information should be evaluated not only in terms of its absolute accuracy but by its closeness to a "perfect theory" or the correct answer to a specific question. This approach is complemented by the work of (BARRY; SCHAMBER, 1998), which highlights the importance of the relevance of information from the users' perspective. Both studies underline the need to consider accuracy and relevance as fundamental criteria in the evaluation.

The model update for the success of information systems proposed by (DELONE;

MCLEAN, 2003) highlights the importance of adapting evaluation criteria to the evolution of technologies and changes in information system management practices. The proposed model offers a robust and adaptable theoretical framework, emphasizing the multiple dimensions of evaluation. Thus, the paper concludes that the success of information systems is a multidimensional and interdependent construct, requiring evaluation of each of its dimensions for a complete understanding. In other words, since QA systems are information systems, their construction must also consider multiple characteristics, including in their evaluation.

The work of (BLOOMA; CHUA; GOH, 2008) provides contributions to the study of QA systems by developing a predictive framework that emphasizes the importance of textual characteristics, such as accuracy, completeness, and reasonableness, in determining the quality of answers. This study, using data from Yahoo! Answers, highlights that the completeness of information is the most impactful factor in users' perception of quality, followed by the reasonableness and accuracy of the answers. This approach corroborates the need for a more granular analysis in evaluating long answers in QA systems, suggesting that besides an answer being accurate, it should meet other criteria, such as completeness.

The study of (STVILIA et al., 2008) about the organization of quality assurance work of information on Wikipedia reveals the complexity of evaluation in large scale collaborative contexts. The research shows that on Wikipedia, the quality of information is directly influenced by criteria such as verifiability and neutrality. The completeness criteria is suggested in the context of the breadth and depth of articles, emphasizing the need to cover all relevant aspects of a topic. In the other hand, the relevance is addressed in the discussion about the importance of keeping the content of articles aligned with Wikipedia's notability criteria, ensuring that the information is pertinent and meaningful to readers. Thus, the study demonstrates how the interaction between automatic assessments and human judgment can help to the maintenance of a high standard of information quality.

In a similar context, the studies of (BLOOMA; CHUA; GOH, 2008) and (STVILIA et al., 2008) highlight the importance of considering the completeness and relevance of the information provided by QA systems. While (BLOOMA; CHUA; GOH, 2008) highlights the completeness of information as an impactful factor in users' quality perception, (STVILIA et al., 2008) shows how verifiability and neutrality are crucial in maintaining the quality of information on Wikipedia, aligning with the notions of completeness and relevance.

The research of (SURYANTO et al., 2009) and the study of (KIM; OH, 2009) address the issue of the relevance of answers in community QA portals. While (SURYANTO et al., 2009) focuses on the importance of selecting relevant answers based on user expertise, (KIM; OH, 2009) identifies a wide range of relevance criteria applied by users, including socio-emotional aspects, highlighting the complexity nature of information evaluation in social environments.

The article by (FICHMAN, 2011) evaluated the quality of answers on QA sites from the perspective of accuracy, completeness, and verifiability, revealing differences among four analyzed platforms: Askville, WikiAnswers, Wikipedia Reference Desk, and Yahoo! Answers. The finding that completeness and verifiability of answers can be improved by multiplying answers, while accuracy is not necessarily.

In (KIM et al., 2017), the importance of relevance, completeness, added value, and web page design as important aspects of information quality that affect the perception of the tourist destination in the tourism websites. Such findings are relevant for understanding how tourist information presented on social media platforms can be optimized to improve the image of tourist destinations in the minds of consumers.

The assessment of the quality and clarity of health information on QA websites is investigated by (CHU et al., 2018), using criteria such as accuracy, completeness, relevance, readability, among others. This study reveals the complexity involved in determining the quality of answers, highlighting the importance of considering individual aspects that contribute to the overall effectiveness of the system. The identification of low-quality answers and the difficulty of users in discerning between high and low quality information show the need for refined metrics that can evaluate more granularly the information provided, thus contributing to a better understanding of the capabilities and limitations of QA systems.

The study by (LI; ZHANG; HE, 2020) offers a perspective on evaluating long answers in academic QA environments. The study shows the importance attributed to criteria such as relevance, completeness, and credibility varies significantly based on specific disciplines, academic positions, and other demographic and contextual factors. In a subsequent study, (LI et al., 2020b) highlights the need for a multidimensional assessment of quality, considering factors that go beyond factual accuracy, that considers the completeness, relevance, and timeliness of information, as well as the authority of the sources for content generated on social networks. In the same context that author analyzed in (LI et al., 2020a) the issue of how researchers judge the quality of answers on academic social QA platforms. The main quality criteria identified include relevance, completeness, and verifiability, with the inclusion of additional criteria such as breadth, the scholarship of the respondent, and the added value of the answers. As result, offering opinions was considered the most important criterion, followed by completeness and added value, highlighting the importance of individual perspectives and analyses in academic answers. Completeness and relevance were especially emphasized as relevant criteria.

The analysis of these studies reinforces the need for a multidimensional evaluation for long answers. The studies offer insights into how different aspects of information and human interaction contribute to the perception of answer quality, highlighting the need for a granular and multidimensional analysis to understand the effectiveness of these systems. These analyzed studies are aimed at analyzing these criteria but do not propose automatic methods of evaluation. The next section presents works related to this thesis that focus on automatic metrics that can be employed in evaluating long answers, mainly those that allow the assessment of the criteria of completeness or relevance.

2.3 Related Works

This section is dedicated to presenting the related works to this research that propose automatic metrics for text evaluation and that allow for the assessment of long answers from QA systems. Moreover, these metrics must have some relation to the specific criteria of completeness or relevance. There are various text evaluation metrics, such as ROUGE (LIN, 2004), BLEU (PAPINENI et al., 2002), and METEOR (BANERJEE; LAVIE, 2005), however, these metrics are not discussed in this section because they serve a general evaluation purpose without a direct relation to the criteria of completeness and relevance. In addition to works proposing metrics, works that contributes with methodologies in conducting metric analyses for text evaluation are considered.

2.3.1 RANKGEN: Improving Text Generation with Large Ranking Models

The work (KRISHNA et al., 2022) introduces RANKGEN, a 1.2-billion-parameter encoder model for English, designed to evaluate model generations from an input sequence (prefix). The main goal of RANKGEN is to consider the limitations of current language models that often assign high probabilities to output sequences that are repetitive, incoherent, or irrelevant to the provided prefix. In other words, these improvements are specifically focused on reducing repetitions, incoherences, and irrelevancies. Thus, RANKGEN aims to enhance the relevance, continuity, and coherence of text generations in relation to the prefix, using large-scale contrastive learning.

RANKGEN operates by calculating the compatibility between a prefix and generations from any pre-trained language model, through the dot product of their vector representations. It is trained using large scale contrastive learning to map a prefix close to the actual continuation sequence that follows it and away from two types of negatives: (1) random sequences from the same document as the prefix and (2) sequences generated from a large language model conditioned on the prefix.

The authors' experiments show that RANKGEN significantly outperforms decoding algorithms such as nucleus, top-k, and typical sampling, both in automatic metrics and in human evaluations with English language writers. The analysis reveals that RANKGEN outputs are more relevant to the prefix and improve continuity and coherence compared to the baselines.

In the context of QA, the prefix can be understood as an input question, while the generated text would be a long answer. Thus, RANKGEN aims to assess the relevance of the answer (generated text) in relation to the question (prefix), and ensure that answer generations maintain logical continuity and cohesion with the provided question. This approach can be seen as a way to assess the relevance of an answer by verifying if the text generation covers relevant aspects related to the prefix (question).

2.3.2 Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary

The work (DEUTSCH; BEDRAX-WEISS; ROTH, 2021) proposes a metric to evaluate the quality of the content of summaries using QA systems. This metric, called QAEval, aims to measure the overlap of information between a candidate summary and a reference summary through pairs of questions and answers. Thus, the proposed model seeks to evaluate the quality of the content of a summary by estimating how much of its content is common with a reference summary. Unlike traditional metrics, such as ROUGE, which are limited to measuring the overlap of tokens in a lexical form or by embeddings, QAEval uses QA pairs to directly measure the information overlap.

The QAEval model works in different steps, starting with the selection of answers in the reference summary, which is used as the basis for generating questions. These questions are then used to interrogate the candidate summary, evaluating the amount of information from the reference summary that is present in the candidate summary. The accuracy of the answers given by the candidate summary is verified, and the final metric is calculated as the proportion of questions answered correctly.

Experiments from the study demonstrate significant correlations with human judgments on benchmark datasets, outperforming or competing with current metrics in different evaluations. Moreover, a detailed analysis of each component of QAEval identified performance bottlenecks, particularly in the QA model and answer verification, indicating areas for future improvements.

The approach used by QAEval is directly related to the criterion of completeness of this research. QAEval uses pairs of questions and answers generated from the reference summary, which could be a long reference answer. By verifying which of these questions can be answered with the generated summary, or the generated long answer, it is possible to check how complete this answer is. Intuitively, it would be possible to calculate the relevance of an answer by creating question and answer pairs from the generated text and checking how many of these questions could be answered with the reference text.

2.3.3 QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization

The work (FABBRI et al., 2022) proposes an optimized metric for evaluating factual consistency in text summarization models. This study aims to analyze the efficacy of metrics based on Entailment (logical implication) and on QA, concluding that the careful selection of the components of a QA-based metric, especially question generation and answerability classification, is crucial for performance. The work contributes with the optimized metric called QAFACTEVAL, which provides a more accurate assessment of factual consistency in summaries generated by Artificial Intelligence (AI) models.

The method employed involves an analysis of the components of QA-based metrics, including answer selection, question generation, answering questions, and evaluating answer overlap. For validation, a comprehensive analysis was conducted using the SumMaC benchmark, which compiles six datasets of factual consistency. QAFACTEVAL demonstrated a significant improvement in performance compared to previous metrics, both QA-based and Entailment-based.

This metric seeks to accurately evaluate factual consistency to ensure that the information provided in a summary is not only present in the source text (completeness) but also relevant to the context of the question or the summary (relevance).

2.3.4 BERTScore: Evaluating Text Generation with BERT

The work (ZHANG* et al., 2020) proposes an automatic metric for text generation evaluation named BERTScore. This metric calculates the similarity of each token in a candidate sentence with each token in a reference sentence, using contextual embeddings to compute token similarities, instead of exact matches. Thus, unlike n-gram based metrics that fail to recognize paraphrases and distant dependencies, BERTScore is more effective at capturing semantic similarity and paraphrasing, making it more robust and better correlated with human judgments.

The metric utilizes embeddings from the BERT model to represent tokens. These embeddings can generate different vector representations for the same word in different contexts, allowing BERTScore to capture the specific use of a token in a sentence. BERTScore computes the similarity of two sentences as the sum of cosine similarities between their token embeddings.

Experiments with BERTScore on automatic translation and image captioning tasks showed a high correlation with human evaluations and superior performance compared to existing metrics like BLEU, METEOR, and ROUGE. BERTScore proved to be robust in a paraphrase adversarial dataset (PAWS), showing greater resistance to challenging examples compared to other metrics.

Correlating BERTScore with the criteria of completeness and relevance, it is important to note that BERTScore provides separate precision and recall metrics, in addition to a combined F1 metric. The precision of BERTScore can be interpreted as an indicator of relevance, as it measures the proportion of information in the system's answer that is relevant to the reference sentence. Meanwhile, BERTScore's recall can be seen as a measure of completeness, evaluating the proportion of relevant information in the reference sentence that is captured by the system's answer.

2.3.5 A Critical Evaluation of Evaluations for Long-form Question Answering

The work (XU et al., 2023) present a methodology focused on the evaluation of long answers generated by QA systems. The main goal of the study is to identify and analyze evaluation methods, both human and automatic, to better understand how long answers should be evaluated in terms of quality. The study aims to identify gaps and challenges in current long answer QA evaluation practices and proposes a more detailed approach to evaluation. Specifically, it suggests that future evaluations move beyond a single “overall score” of the answer and adopt a multi-feature evaluation, focusing on aspects such as factuality and completeness.

The research analyzed both human evaluations and automatic text generation metrics to assess long QA answers. For human evaluations, domain experts were hired to judge pairs of answers based on detailed criteria, such as the completeness of the answer. For automatic evaluations, a series of 12 automatic metrics was examined to determine their correlation with human judgments, especially in fine aspects such as coherence and fidelity of the answers.

The study concluded that none of the existing automatic metrics is predictive of human preference judgments regarding the overall quality of the answers. However, some automatic metrics show potential in modeling more detailed aspects of the answers, which may stimulate research on a new generation of automatic metrics for long answer QA systems. The study’s results highlight the importance of considering completeness and factuality as decisive criteria in evaluations. Domain experts valued these aspects when preferring one answer over another. This finding suggests the need for evaluation metrics that can adequately capture the completeness and accuracy of the information provided in long answers.

2.3.6 Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text

The work (GEHRMANN; CLARK; SELLAM, 2023) provides a critical analysis of evaluation practices in Natural Language Generation (NLG) and suggests future directions for research and improvements, identifying weaknesses in evaluation practices and proposing solid steps to improve the accuracy and reliability of these evaluations. The work emphasizes the need to provide a realistic view of the models’ limitations, through the creation of evaluation reports that detail the models' limitations instead of just showing high performance numbers. It is also suggested to adopt complementary sets of automatic metrics, rigorous human evaluations, and the release of data that allow for reanalyze with improved metrics.

A highlighted result from the work is the insufficiency of automatic metrics to evaluate focusing on multiple characteristics. According to the authors, currently, automatic metrics primarily focus on the lexical similarity between the model output and human references, which may not capture well the quality of the generated content or its appropriateness to the context.

2.3.7 General Analysis of Related Works

The related works reveal a possible diversification in the evaluation of long answers generated by QA systems, with various approaches being explored to measure the quality of the generated text, mainly considering the criteria of completeness and relevance. The studies contribute to the understanding and improvement of evaluation metrics, highlighting the complexity of evaluating long answers and the need for individual metrics to capture different characteristics.

In summary, the analysis of related works underlines the need to develop and refine individual evaluation metrics for long answers from QA systems. These metrics should be capable of capturing not only the correlation between the evaluated text and a reference text but also the relevance of information in relation to all the information provided in the answer, as well as the completeness of this answer. The search for such metrics is important for advancing the development of QA systems that can generate answers that are not only accurate but also complete and contextually relevant, effectively meeting the informational needs of users.

3 SYSTEMATIC REVIEW OF NON-FACTOID QA

This chapter present a systematic review focuses on the state-of-the-art for non-factoid QA systems, observing different tasks and methods, as well as the available databases, evaluation strategies, outcomes, and recommendations for future research. Therefore, the objective of the review is to answer the following research questions:

1. What are the methods and tasks involved in non-factoid QA systems?
2. What are the data sets available for non-factoid QA systems?
3. What are the limitations of non-factoid QA?

A Systematic Review differs from traditional narrative reviews by adopting a replicable, scientific, and transparent process to minimize bias through exhaustive literature searches (TRANFIELD; DENYER; SMART, 2003; DENYER; TRANFIELD, 2009; HIGGINS et al., 2019). Despite the relative maturity of systematic reviews, there is no firm agreement about the number of stages for conducting a systematic review. For example, while the Cochrane Reviewers' Handbook (HIGGINS et al., 2019) and National Health Service Dissemination (2001) agreed in 9 Stages for a systematic review, recent studies have used a simplified approach (KHAN et al., 2003; DYBå; DINGSøYR, 2008). We have adopted the same method proposed by Dybå (DYBå; DINGSøYR, 2008) by breaking down the study into 6 stages.

3.1 Protocol

The protocol is a document that gives a general overview of how the review is performed. Typically, it specifies the research questions, search strategy, inclusion, exclusion and quality criteria, data extraction, synthesis method, etc. In previous studies, we found different guidelines for designing a protocol, which complements each other. Therefore, we have relied on guidelines presented in (TRANFIELD; DENYER; SMART, 2003; DENYER; TRANFIELD, 2009; HIGGINS et al., 2019) for designing our protocol.

3.2 Systematic Search

A systematic search begins with the definition of search terms and electronic databases to be scrutinized. The search terms are used in databases of scientific articles (e.g., Web of Science or Google Scholar) for retrieving only related work, in our case on non-factoid QA systems. This systematic review particularly addressed to non-factoid QA systems; Consequently, we had to create our own set of terms and electronic databases built based on related studies (DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019; KOLOMIYETS; MOENS, 2011), and fine-tuned during discussions with the review team composed by the authors of the paper (CORTES et al., 2022). The final set of search terms is presented below:

1. non-factoid;
2. definition question;
3. confirmation question;
4. causal question;
5. comparative question;
6. opinionated question.

In order to select studies that respect both criteria, the keywords were combined through the Boolean “OR” and “AND” operators to formulate a query string, as following:

QUERY STRING = (1 OR 2 OR 3 OR 4 OR 5 OR 6) AND “question answering"

Only papers that contain the patterns described in the query string were considered in this review. The following electronic databases were employed in this study:

DATABASES = [*“ACM Digital Library", “Web of Science", “IEEE Xplore", “Science Direct - Elsevier", “Springer Link"*]

3.3 Inclusion and exclusion criteria

We have developed different inclusion and exclusion criteria to reduce the number of unrelated and less significant papers for the manual review. For instance, to retrieve only updated and relevant works, we have only considered studies written in English since 2010 published in international conferences. Additionally, we have performed a semi-automatized analyzes using a tool to make sure only papers focusing on the QA topic were considered. We employed Rayyan QCRI (OUZZANI et al., 2016), which is a tool used to highlight a set of keywords in the papers and facilitate the process of inclusion and exclusion. When a paper present several keywords highlights, it is assumed that the paper covers the review's topics and should be included. Furthermore, once our research focused on non-factoid questions, we excluded studies that employ only factoid questions in their experiments or do not propose any method for non-factoid questions. We also excluded surveys and reviews from our study. In sum, the paper is included in our study only if it fulfills all the following criteria:

- Written in the English language;
- Published between 2010 and 2023;
- Focus on QA systems, as indicated by the search terms;
- Consider the challenges of non-factoid questions in the solutions;
- Employ non-factoid questions in the experiments.

3.4 Work Eligibility and Quality Control

In order to ensure that only relevant studies are included in this systematic review, we have created a quality control system, which consists of questionnaire. To answer the questionnaire we used three annotators – the authors of this review – to read the papers and answer a questionnaire regarding a paper's eligibility and quality. The questionnaire contains questions with three possible options: “Yes", “Partial" and “No". Only studies which have “Yes" for most of the questions and none “No" were considered. The questions are:

- Does the study address empirical research, or is it a merely “lessons learned” report based on an expert's opinion?
- Were the objectives and conclusions clearly reported?
- Is not the study an example of editorials, prefaces, article summary, interview, new, or review?
- Does the study provide an understandable description of the proposed methods?
- Was the research methodology suitable for the aims of the research?
- Was there a description of the data sets employed, and whether they contain non-

factoid questions?

- Does the data set present quality data, enough instances, and was it adequate for the experiments?
- Does the study employ adequate methods to analyze the data?
- Were the results calculated using metrics appropriate to the experiment?

3.5 Data Extraction and Annotation Process

To answer the research questions posed in this study, we have employed the software Rayyan QCR for a semi-automatic annotation of papers. The software enables the extraction of metadata, such as author, institution, year, etc. Simultaneously, the annotator were responsible for extracting in-depth information which was not explicit on the paper, such as the name of the methods or the data sets employed. We used three annotators, being at least two annotators per work, and a third one doing disambiguation for the cases where the two annotations did not agree on a particular label. In sum, we have annotated the following set of features from the papers:

1. **Year of publication:** the year that the study was published;
2. **Language:** these are the languages of the data used in the analyzed study;
3. **Question type:** the types of question used in the study experiments. We consider a taxonomy of question types derived from (DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019) that consist of a) *Definition*: questions requiring a definition; b) *How*: questions requiring an instruction; c) *Why*: questions requiring a reason; d) *Opinion*: questions requiring an opinion; e) *Comparison*: questions requiring a comparison between entities; f) *Conformation*: questions checking a fact; g) *Factoid Included*: the data set include factoid questions;
4. **Domain of knowledge:** The knowledge is the proposed method has focused on. We first classify the study into the open-domain or the closed-domain. When classified in the closed-domain, we also identify which knowledge area it is focusing on;
5. **Knowledge source type:** this feature corresponds to the characteristics of the information source used by the QA system. We propose four categories derived from the reviews (SOARES; PARREIRAS, 2020; DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019): *Documents*: the system uses a collection of raw text documents; *Web*: the system uses information retrieved from the web, such as pages

resulting from a search engine; *Knowledge Graph*: uses a structured knowledge base; *Answer List*: usually used by community QA systems, in which for each question, the system consults a list of possible candidate answers;

6. **Metrics used for Evaluation:** The evaluation metrics used by the study to evaluate the proposed methods;
7. **Research problem:** the QA tasks the study has focused on. We classify each problem focused by the study on one of the tasks presented in Section 2.1. Thus, the following tasks are associated with the study analyzed when it follows the criteria:
   - *Question Classification*: assigned to studies with methods related to text classification that somehow classify the input question. It includes tasks like answer type classification and topic classification;
   - *Question Reformulation*: assigned to studies that enhance the question by transforming it into semantically equivalent, like improving the question text with data from the WordNet;
   - *Document Retrieval*: assigned to studies that aim to retrieve documents with relevant information, like text documents and web pages.
   - *Passage Extraction*: assigned to studies that aim to provide methods to extract passages from the retrieved documents;
   - *Candidate Answer Extraction*: assigned to studies that propose methods to identify answer candidates in a list of text passages;
   - *Candidate Answer Ranking*: assigned to studies that propose methods to rank or select the candidate answers most likely to be correct;
   - *Answer Generation*: assigned to studies that propose composing a final answer using the information extracted from previous tasks. It includes methods like NLG and summarization;
8. **Method employed:** the methods proposed by the study to solve each task;
9. **Dataset employed:** the data collections used in the experiments;
10. **Results:** results and conclusion of the study.

3.6 Synthesis of the findings

The information extracted from the papers in this review were summarized in a tabular format, where each row represents a study and each column represents an extracted feature. The tabular organization enables comparison across works, and reciprocal translation of findings into a higher-order of interpretation, as well as it is a well-employed method and highly recommended for qualitative data analysis (SEERS, 2012; CORBIN; STRAUSS, 2014).

3.7 Limitations of this review

The main limitation of this review is related to a possible bias in the selection of the studies. Nevertheless, we aimed at reducing this bias by querying the digital databases following the standards and keywords used by previously systematic reviews. Another possible limitation is related to coverage since we only covered studies focused on QA systems that explicitly performed experiments addressing non-factoid questions. Consequently, related problems such as automatic text summarization were not included because they were not evaluated using QA benchmarks and standards.

3.8 Review Results

Our systematic review initially covered a total of 455 studies; nevertheless, after carefully removing duplicates, this number reduced to 698 papers. The inclusion and exclusion criteria were divided into two steps: while the first step removed 165 unrelated studies based on titles and keywords, the second step excluded 201 studies based on a manual reading of the abstract. During the eligibility and quality control, we employed semi-automatic annotation to remove 48 studies that did not fulfill the quality criteria. Therefore, we considered only 125 studies for a manual analysis listed in Appendix A and Appendix B. Figure 3.1 summarises the systematic review process.

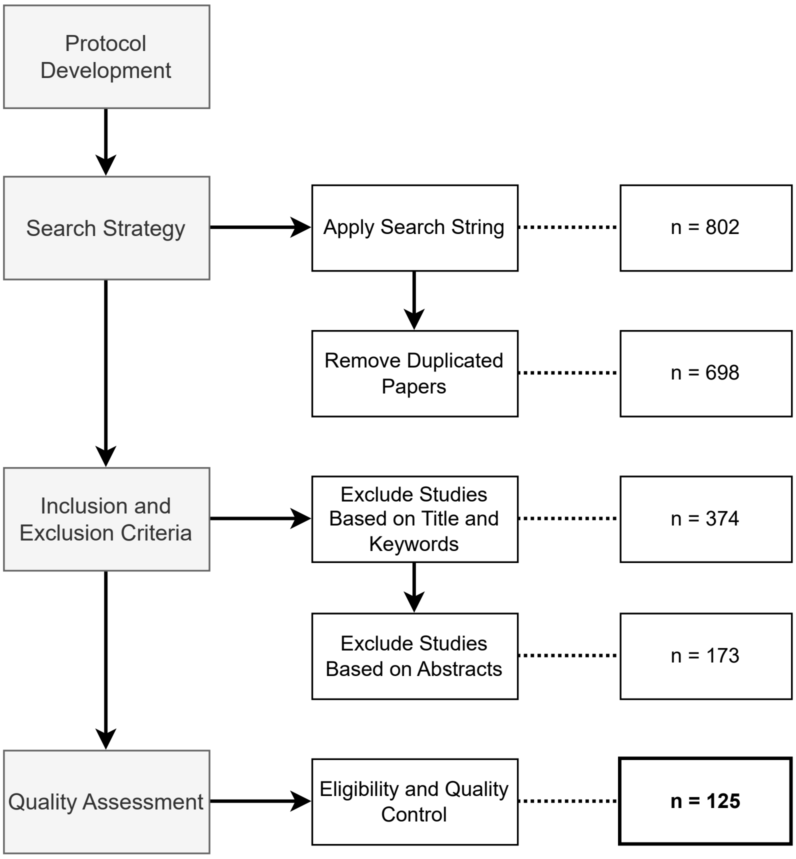

Protocol Development
Search Strategy
Apply Search String
n = 802
Remove Duplicated Papers
n = 698
Inclusion and Exclusion Criteria
Exclude Studies Based on Title and Keywords
n = 374
Exclude Studies Based on Abstracts
n = 173
Quality Assessment
Eligibility and Quality Control
n = 125

Figure 3.1 – Systematic review processes.
Source: The Author.

3.8.1 Publication over the years

In our study, we have observed a recent interest in non-factoid QA. Figure 3.2 presents the works' distribution over the years. It shows a reduction of publications in 2010 and a rise in interest since 2015. Specifically, the number of studies began to increase gradually after 2015, suggesting the beginning of a more focused investigation into non-factoid QA systems. This initial growth in research interest may be associated with advances in machine learning and NLP that facilitated more sophisticated approaches to QA systems.

There is a considerable increase from 2021 onwards, with the number of published papers peaking in 2022. This peak can be related to the widespread adoption of new language models, particularly those based on Transformer architectures such as BERT. These models have significantly improved the ability to understand and generate text, thus enhancing the performance of non-factoid QA systems.

The year 2023 contains a smaller number of articles, but this figure is expected to rise as more articles are indexed. The data collection was carried out at the beginning

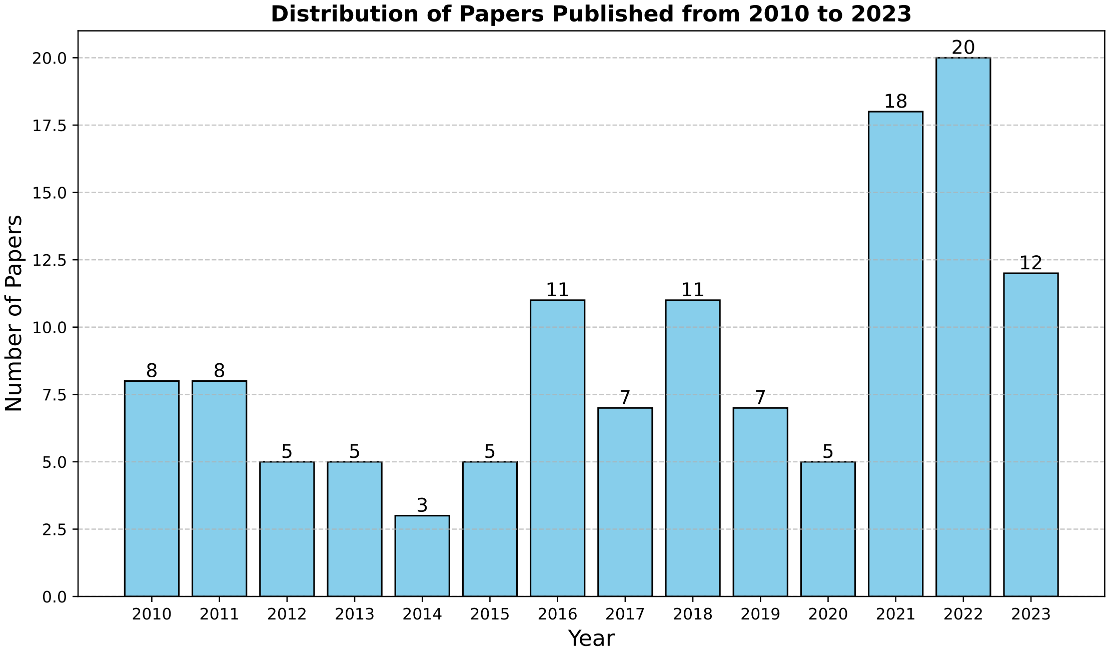


Figure 3.2 – Number of publication over the years.
Source: The Author.

of 2024, and thus some publications from 2023 may not have been included at that time. The preliminary data indicates a continued interest in non-factoid QA research, reflecting the ongoing advancements in language models and their applications.

Also, about 20% of the studies were published in journals, while the leftover was published in the conference literature. We observe that most of the analyzed studies focus on specific tasks and techniques of a QA system and not on the system as a whole. Thus, these studies often do not have enough content to justify their publication in a journal, fitting better into the conference literature.

3.8.2 Language Focus

Concerning Language, as expected, non-factoid QA systems have mainly addressed English documents. Among all studies, 73.6% (92 studies) used English data in the experiments, and only 3.2% (4) use a multi-language strategy (P26, P27, P48, P54). Among non-English works, Chinese is the most addressed Language representing a total 9.6% (12) (P8, P26, P27, P36, P42, P48, P65, P66, P73, P76, P116, P117), followed by Arabic with 7.2% (9) (P35, P48, P51, P57, P72, P81, P89, P95, P118), and Japanese with 6.4% (8) (P5, P50, P55, P56, P59, P92, P97, P105). Other less used languages were Indonesian with 2.4% (3) (P34, P106 and P109), Persian with 1.6 % (2) (P88 and P112), Italian (P54), Korean (P49), and Russian (P3).

3.8.3 Types of Non-factoid Questions

Most of the studies (40%) addressed definition (what), casualty (how), and reasoning (why) questions (P1, P5, P8, P9, P10), followed by studies focused on comparison and confirmation questions (14.4%) (P2, P3, P4, P7, P11). Table **??** gives an overview of the different types of questions addressed by non-factoid QA works. We observed that studies addressing definition, casualty, or reasoning usually address multiple types of questions because of their similarity. Nevertheless, we found many studies have addressed a single type of question, such as confirmation or comparison. Although most of the studies have focused on non-factoid questions, several works (24%) also included factoid questions in their experiments (P3, P4, P19, P20, P33).

Question Type	Studies
Definition	18
How	17
Why	15
Comparison	10
Confirmation	8
Opinion	7
Factoid Included	30

*The sum of studies is less than the total of analyzed paper because we just included papers that specified the question type.

Table 3.1 – Distribution of studies in relation to the type of question.
Source: The Author.

3.8.4 Application Domain

Many studies focused on an open-domain, such as the class of questions where the answer is found on *Wikipedia* or *Yahoo! answers*. In total, 78.4% (98) of studies focused on an open-domain, while only 21.6% (27) focused on a closed-domain, such as health and insurance. Although, the main areas of application among the closed-domains are health and insurance, we have also observed that non-factoid QA systems have been used for answering questions in the domain of agriculture (P27), biology (P47), E-commerce (P12), financial (P18), geography (P8), political (P51), tourism (P36), environment (P76), patents (P78), religion (P88), cooking (P91), education (P93, P95), science (P109), food safety (P116) and law (P123).

3.8.5 Knowledge Source

Most of the works use unstructured data such as textual documents (P5, P7, P23) and webpages (P16, P28, P59) for answering questions. Table 3.3 presents the knowledge source type used by the analyzed studies. In total, 56% (72) of works use textual documents, 23.2% (29) use webpages, 17.6% (22) use a list of possible answers, and 11.2% (14) use knowledge graphs as a source of knowledge. Few studies employ a combination of different knowledge sources, such as knowledge graphs and documents (P8, P12, P62).

Table 3.2 – Studies distribution over the type of Knowledge Source.

Knowledge Source	Studies
Documents	72
Web	29
Answer List	22
Knowledge Graph	14

Table 3.3 – Studies distribution over the type of Knowledge Source.
Source: The Author.

3.8.6 Most Employed Metrics for Quality Assessment

Before, we present a brief explanation of the most common metrics and evaluation methods used by the analyzed studies.

- **Mean Reciprocal Rank (MRR)** is a statistic measure for evaluating any process that produces a list of possible answers to a sample of queries, ordered by probability of correctness. Regarding QA systems, giving a set of questions Q, the Mean Reciprocal Rank is definied as:

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

  where $rank_i$ is the rank position of the first relevant answer for the i-th question (DIMITRAKIS; SGONTZOS; TZITZIKAS, 2019).

- **Precision@k (P@k)** corresponds to the number of relevant results among the top-k-answers. For example, the precision of a QA model that return k possible answers for a question q is given by:

$$P@k = \frac{|Found(q)|}{k}$$

where $Found(q)$ is the list of correct answers returned by the model.

- **Mean Average Precision (MAP)** evaluates the mean of the average precision for a set of queries. It is mainly used to evaluate ranked-results. For example, giving a set of questions Q, the MAP is calculated by:

$$\frac{MAP = \sum_{q=1}^{|Q|} AveP(q)}{|Q|}$$

where $AveP(q)$ consider the order in which the returned result are presented by computing a precision and recall at every position in the ranked sequence of results. Therefore, it is the average value of the precision as a function of the recall of the question q.

- **Accuracy** is the fraction of the questions that are answered correctly. For example, for a set of questions Q, the Accuracy is calculated as:

$$Accuracy = \frac{|CQ|}{|Q|}$$

where CQ are those questions that were answered correctly.

- **F-Score** is a weighted harmonic mean between precision and recall. The F-Score is computed as:

$$F_\beta = \frac{(1+\beta^2) \cdot (precision \cdot recall)}{\beta^2 \cdot precision + recall}$$

.

- **Normalized Discounted Cumulative Gain (nDCG)** measures the usefulness, or gain, of a document based on its position in the result list. It is normally employed to measure of ranking quality. The nDCG is computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where p is the particular rank position. The DCG_p is the discounted cumulative gain and penalizes highly relevant answers that appear lower in the rank of answers candidates. It is computed as:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i_1)}$$

where rel_i is the graded relevance of the result at position i. The $IDCG_p$ is computed as:

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{log_2(i_1)}$$

where REL_p is the the list of relevant answers ordered by their relevance.

- **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** evaluates the answers returned by the QA system, comparing them against correct answers. It works by comparing the produced answer against a set of reference answers. The $ROUGE_{Recall}$ is computed as:

  $$ROUGE_{Recall} = \frac{overlaps}{correct\_words}$$

  where $overlaps$ is the number of overlapping words between the produced answer and the correct one. $correct\_words$ is the total of words in the correct answer. The $ROUGE_{Precision}$ is computed as:

  $$ROUGE_{Precision} = \frac{overlaps}{produced\_words}$$

  where $produced\_words$ is the total of words in the produced answer.
  There are different variants of ROUGE. For example, ROUGE-N, which considers the overlap of N-grams, and ROUGE-L evaluates the longest common subsequence between the predicted and correct answer.

- **Human Assessment** is an evaluation strategy mainly used for real-time competitions, such as LiveQA Trec (AGICHTEIN et al., 2015). It uses humans to evaluate QA systems through a manual judgment of answers.

Table **??** presents the most used metrics to evaluate QA systems. The most employed metrics are MRR (39.2%), F-score (35.2%), and P@K (25.6%), used mainly for assessment of Candidate Answer Extraction (P4, P13, P17) and Candidate Answer Ranking (P20, P21, P22). Besides, some studies have employed Accuracy (20%) to assess Question Classification, Candidate Answer Extraction and Candidate Answer Ranking

Metric	**Associated Task**	**Studies**
MRR	Question Reformulation Document Retrieval Passage Extraction Candidate Answer Extraction Candidate Answer Ranking	49
P@K	Candidate Answer Extraction Candidate Answer Ranking	32
MAP	Candidate Answer Ranking	26
Accuracy	Question Classification Candidate Answer Extraction Candidate Answer Ranking	25
F-score	Question Classification Question Reformulation Answer Generation	44
NDCG	Candidate Answer Extraction Candidate Answer Ranking	16
ROUGE	Answer Generation	13
Human Assessment	Candidate Answer Ranking	8
Others		21

Table 3.4 – Evaluation strategy used by studies.
Source: The Author.

(P48, P62, P74). The method called “Human Assessment" (6.4%) represents a manual evaluations – mainly used for competitions, such as LiveQA Trec –, where experts assess the systems in real-time (AGICHTEIN et al., 2015).

3.8.7 Tasks involved and Methods Used for Non-factoid Question Answering Systems

We divided our analysis according to the task present in Section 2.1. We observed that 81.6% (102) of the works have focused on Answer Processing, followed by 24.8% (31) on Question Processing, and 12.8% (16) on Information Retrieval. Many studies have focused on Answer Processing since this component is different from the conventional factoid QA system. Many studies have also focused on Question Processing, which shows concern for extracting pertinent information for non-factoid questions. Table **??** summarizes the distribution of the tasks over the works.

Candidate Ranking is the most addressed task by the analyzed studies. The principal strategy employed in these studies is supervised learning (P27, P36), where the goal is to rank a list of potential answers. There are two main approaches, namely ranking and classification. Regarding ranking, most of the works try to estimate the distance between

Architecture Stage	Task	Studies
Question Processing (31)	Question Classification	21
	Question Reformulation	11
Information Retrieval (16)	Document Retrieval	9
	Passage Extraction	9
Answer Processing (102)	Candidate Answer Extraction	50
	Candidate Answer Ranking	72
	Answer Generation	15

*The sum of studies go over the total of analyzed papers, once it is possible to assign more than one stage or task by study.

Table 3.5 – Studies distribution over QA architecture stages and tasks.
Source: The Author.

question input and the answer candidate. While some works set a score for each candidate based on lexical and semantic features related to the difference between the input question and the answer candidate (P23, P39, P49), other studies propose learning to rank strategies based on lexical, semantic, and other textual features extracted from the text using machine learning models (P19, P20, P54). Conversely, some studies treated this task as a binary classification problem, where the system classifies the candidate answer as correct or incorrect (P15, P27, P67). Not least of all, some studies presented considerable results using pre-trained models, and attention mechanisms, such as BERT (P1, P5). These studies show that these methods based on neural model significantly outperform other models, such as BM25, an effective term-matching retrieval model.

The second most addressed task in non-factoid QA systems is *Candidate Answer Extraction*. The most employed approaches are based on traditional information retrieval methods, such as BM25, which estimates the relevance of a document giving a query (P17, P19), to deep neural models (P13, P16, P23). These studies show that deep neural models, such as Long short-term memory (LSTM) and Convolutional Neural network (CNN), present better results than traditional ones. Few studies employ summarization methods to create candidate answers (P16, P29, P63), such as deep auto-encoder and LSTM auto-encoder for sentence representation. Some other studies also use answer lists from community QA websites selecting text fragments that are more likely to bear answers to the query. Experiments show a positive impact on the performance optimization-based summaries (P23). Furthermore, several studies have used knowledge graphs to support Candidate Answer Extraction (P2, P4, P15, P66) by harness the unique properties of knowledge graphs to treat data redundancy, access the links between data objects, run efficient queries against the knowledge graph and explore the updated nature of the knowledge. Also, some studies employ metathesaurus and sentiment analysis for answer extraction (P4).

Works addressing *Question Classification* have combined different features, such as lexico-syntactic (P42, P62) and sentiment analysis (P58). Handcrafted rules usually perform well due to experts' effort to create manual rules (CORTES et al., 2020), however only few works have proposed a combination of handcrafted rules with lexico-syntactic patterns to classify questions (P3, P4). We have also observed studies proposing new taxonomy for non-factoid questions that best fit specific domains (P51, P61). The results suggest that new taxonomy with multi-label classification is better than a single-label, once it helps to reduce the search space for answers.

Studies on *Question Reformulation* have proposed methods based on the extraction of different information from the questions. For example, (P8, P15) decompose the question into sub-queries and resolve each of them individually to create a final answer. On the other hand, some studies have expanded the question using external knowledge bases, such as Wikipedia or a knowledge graph (P64, P71).

The studies on *Document Retrieval* have directly looked for the answers on web pages using search engines (P19, P28, P64). Usually, these studies use commercial search engines, such as Google Search API and Bing, to mine answers candidates from the top web pages retrieved. Also, some studies use the search engines themselves to expand the questions using the snippets of the top search results (P64). Few studies have proposed methods for dealing with technical terminology of particular domains through special encoders (P4, P7) applying metathesaurus, synonyms, and cross-attention mechanisms between the query and document words to discover the important terms.

In spite of *Answer Generation*, we observed that most studies preferred to use multiple passages extracted from the original document instead of generating a single answer using NLG. The few works addressing *Answer Generation* proposed end-to-end methods that extract context information from documents to generate the answer with neural models (P40, P41, P78, P87, P88, P92, P99, P101, P103, P105 and P110). Some of them also employed external knowledge to capture deep semantic relationships between sentences and questions to acquire question-aware representations for the document (P40, P41).

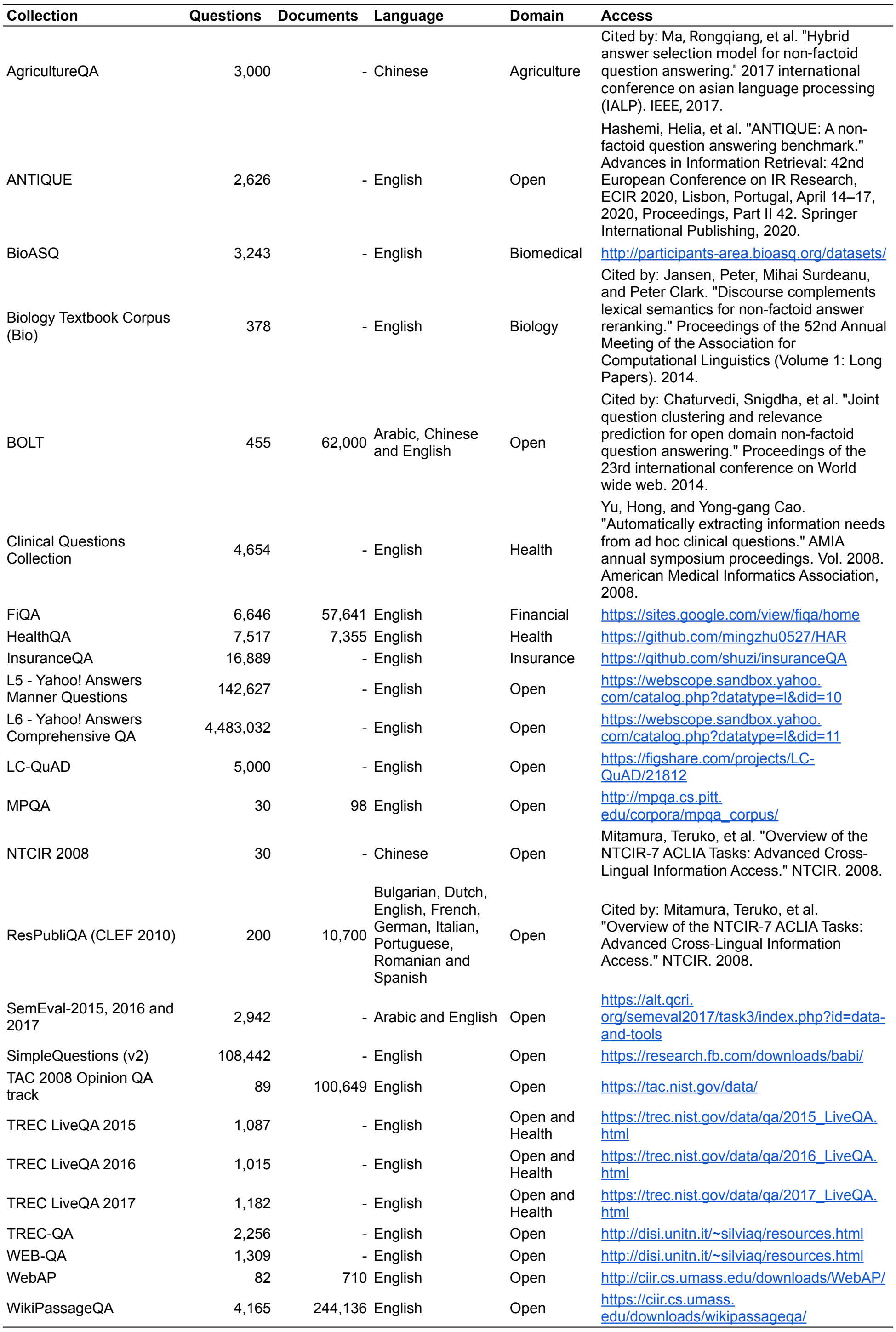
3.8.8 Data sets

We have observed that many works do not use standard benchmarks for evaluating their systems. While some of them use a subset of an existing data set, others build a new data set from scratch. During the analysis process, we have noticed that the works tend to modify the data sets according to their experiments' needs. Therefore, we have found several different versions of the same data set and overlapping of data. Table 3.6 shows those data sets and some of their relevant features for the QA research.

Among the data sets used, several can be classified as community QA data sets. They are collections composed of questions and answers created by users from community QA web portals. The majority of a community QA data set questions are not trivial to be answered with a simple web search. Therefore these questions can be classified as complex and, most of the time, as non-factoid. Also, different from conventional collections that are created requiring the user to invent a question for given information, this type of collection has the advantage of being naturally created by the user in natural conditions of questioning.

The main difference between the Community QA data set and the conventional ones is that the Community QA collections' answers should be considered candidate answers (BAE; KO, 2019; KHUSHHAL et al., 2020). It is usually not validated by certified experts and has a score or a ranked order from users' votes. Also, (SURDEANU; CIARAMITA; ZARAGOZA, 2008; YAN; ZHOU, 2015) describe that these candidate answers have a high variance of quality like answers range from exceptionally informative to completely irrelevant, and someones can be even abusive.

Unlike conventional End-to-End QA, the task involving Community QA data sets usually relies on selecting the most appropriate answers from a given list of answers candidates for the target questions. The author usually picked the most voted answer as the correct candidate to elaborate on the list of candidate answers during these data sets' construction. The rest of the irrelevant one is picked from answers candidates of other questions (COHEN; YANG; CROFT, 2018). Therefore, most of the candidate answers list may not have any semantic relationship to the target question.



Collection	Questions	Documents	Language	Domain	Access
AgricultureQA	3,000	-	Chinese	Agriculture	Cited by: Ma, Rongqiang, et al. "Hybrid answer selection model for non-factoid question answering." 2017 international conference on asian language processing (IALP). IEEE, 2017.
ANTIQUE	2,626	-	English	Open	Hashemi, Helia, et al. "ANTIQUE: A non-factoid question answering benchmark." Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42. Springer International Publishing, 2020.
BioASQ	3,243	-	English	Biomedical	http://participants-area.bioasq.org/datasets/
Biology Textbook Corpus (Bio)	378	-	English	Biology	Cited by: Jansen, Peter, Mihai Surdeanu, and Peter Clark. "Discourse complements lexical semantics for non-factoid answer reranking." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.
BOLT	455	62,000	Arabic, Chinese and English	Open	Cited by: Chaturvedi, Snigdha, et al. "Joint question clustering and relevance prediction for open domain non-factoid question answering." Proceedings of the 23rd international conference on World wide web. 2014.
Clinical Questions Collection	4,654	-	English	Health	Yu, Hong, and Yong-gang Cao. "Automatically extracting information needs from ad hoc clinical questions." AMIA annual symposium proceedings. Vol. 2008. American Medical Informatics Association, 2008.
FiQA	6,646	57,641	English	Financial	https://sites.google.com/view/fiqa/home
HealthQA	7,517	7,355	English	Health	https://github.com/mingzhu0527/HAR
InsuranceQA	16,889	-	English	Insurance	https://github.com/shuzi/insuranceQA
L5 - Yahoo! Answers Manner Questions	142,627	-	English	Open	https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=10
L6 - Yahoo! Answers Comprehensive QA	4,483,032	-	English	Open	https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11
LC-QuAD	5,000	-	English	Open	https://figshare.com/projects/LC-QuAD/21812
MPQA	30	98	English	Open	http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
NTCIR 2008	30	-	Chinese	Open	Mitamura, Teruko, et al. "Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access." NTCIR. 2008.
ResPubliQA (CLEF 2010)	200	10,700	Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish	Open	Cited by: Mitamura, Teruko, et al. "Overview of the NTCIR-7 ACLIA Tasks: Advanced Cross-Lingual Information Access." NTCIR. 2008.
SemEval-2015, 2016 and 2017	2,942	-	Arabic and English	Open	https://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools
SimpleQuestions (v2)	108,442	-	English	Open	https://research.fb.com/downloads/babi/
TAC 2008 Opinion QA track	89	100,649	English	Open	https://tac.nist.gov/data/
TREC LiveQA 2015	1,087	-	English	Open and Health	https://trec.nist.gov/data/qa/2015_LiveQA.html
TREC LiveQA 2016	1,015	-	English	Open and Health	https://trec.nist.gov/data/qa/2016_LiveQA.html
TREC LiveQA 2017	1,182	-	English	Open and Health	https://trec.nist.gov/data/qa/2017_LiveQA.html
TREC-QA	2,256	-	English	Open	http://disi.unitn.it/~silviaq/resources.html
WEB-QA	1,309	-	English	Open	http://disi.unitn.it/~silviaq/resources.html
WebAP	82	710	English	Open	http://ciir.cs.umass.edu/downloads/WebAP/
WikiPassageQA	4,165	244,136	English	Open	https://ciir.cs.umass.edu/downloads/wikipassageqa/

Table 3.6 – Data sets used by the studies.
Source: The Author.

3.8.9 Limitations of Non-factoid Question Answering

The limitation of the analyzed studies is related to how they provide the answer. The majority of studies focus on selecting few passages from different documents and ranking them according to their usefulness to answer a question. However, it is common for non-factoid questions to have several restrictions, narrowing the search space down to a specific answer. For instance, for the question “How should I treat measles in a 12-year-old boy?" the ideal passage to be used as an answer should cover “treatment", “measles", “12-year-old" and “boy", which is very unlikely and there may not be a ready-made passage in the knowledge base containing all the information needed. In this case, the ideal system must search for different information pieces in different documents and merge them to compose a single answer. However, this is challenging and still an open research problem. Some works have tried to overcome this limitation by presenting a set of sentences grouped by the terms (P4, P60, P70). However, this approach still requires a great interpretation effort from the user.

Regarding evaluation, non-factoid QA requires a great deal of manual effort to verify the system’s correctness. Unlike factoid QA systems, where a question usually has one of few correct alternatives, answers for non-factoid questions can be expressed in infinitive manners. Therefore, it is challenging for humans to assess end-to-end non-factoid QA systems.

3.8.10 Systematic Review Conclusion

We presented a systematic review of the literature addressing non-factoid QA systems. From a total of 455 recent studies, we selected 75 papers based on our quality control system and exclusion criteria for an in-depth analysis. This review aims to explain the particular aspects of non-factoid QA systems, such as the distinct tasks and methods, the available benchmarks, and the different types of questions addressed in recent works. This systematic review helped to answer the following questions:

What are the tasks and methods involved in non-factoid QA systems? We observed that the general architecture of non-factoid QA systems does not differ from factoids. Nevertheless, the methods employed in each task of the non-factoid QA system vary to some extent. For example, our empirical analysis showed that many studies on non-factoid questions have focused on *Candidate Answer Extraction*. While in factoid question, the *Candidate Answer Extraction* is responsible for extracting entities as a possible answer candidate, non-factoid QA systems extract multiples passages from a document(s) to compose a single answer. The methods used in this task vary from BM25 based methods (P1, P4, P28, P60) to Deep Neural models (P12, P17, P22, P43). Our review also revealed that although the composition of an answer based on multiple passages is one of the most distinct characteristics of non-factoid questions, only a few works have addressed this problem so far. The only few works that have addressed this issue have used Automatic Text Summarization to digest multiple passages retrieved by previous steps and generate the user's final answer (P16, P29, P63).

What are the data sets available for non-factoid QA systems? We have found an increasing number of available data sets for non-factoid questions. As expected, most data sets address the English Language; however, we have found data sets for non-English Languages such as Chinese, Arabic, and Japanese. Regarding area of application, most of the available data sets addresses health and insurance, followed by agriculture (P27), biology (P47), E-commerce (P12), financial (P18), geography (P8), political (P51), and tourism (P36). We have also noticed that many works tend to modify the data sets according to their needs. Consequently, different versions of the same data set are used for evaluation, which makes a fair comparison between the studies difficult.

What are the limitations of non-factoid QA? Automatic generation of answers based on multiple passages is a critical issue for developing full end-to-end non-factoid question-answer systems. The problem emerges from the fact that automatic generation of coherent and cohesive text – especially for long passages – is still an open research question (BROWN et al., 2020). Broadly speaking, coherence and cohesion refer to how a text is organized so that it can hold together. In a coherent answer, concepts are connected meaningfully and logically by using grammatical and lexical cohesive devices. Furthermore, evaluation of an end-to-end non-factoid QA system seems to be a challenging issue. Although quality estimation is a critical component for developing better systems, this kind of problem is not exclusively of QA systems but also for all text-to-text applications, such as machine translation, text simplification, text summarization, grammatical error correction, and NLG (SPECIA; SCARTON; PAETZOLD, 2018).

The future directions in the non-factoid QA should concern methods that generate natural language and use several information sources to compose complex answers instead of using a simple extracted sentence. Also, there are challenges in other QA pipeline stages to compose answers through structured knowledge bases and extract relevant information from complex questions. Finally, researchers must seek to share the same data set versions in their experiments to compare results between proposed methods.

4 METHODOLOGY

This chapter presents the research methodology used in this thesis, considering all the research stages carried out and their relationships. Figure 4.1 provides a summarized flow of these stages. Throughout the chapter, the focus is mainly on the detailed steps of the methodology for evaluating the proposed metrics through a benchmark created for this purpose. In general, the goal is to assess how similar the evaluated metrics are to human evaluation. Therefore, the research stages can be briefly summarized as follows:

1. Create a set of long answers evaluated by humans in terms of their completeness and relevance;
2. Apply the evaluated metrics to each answer in the set;
3. Check the similarity between the evaluated metrics and the human annotation;

Thus, the closer a metric's similarity to human annotation, the better its performance in determining the relevance or completeness of an answer. Figure 4.1 presents the schema of the methodology, that started with a systematic review of non-factoid QA, followed by the development of a benchmark for experiments. This benchmark includes the creation of a dataset, the selection of baseline metrics, the deleopment of proposed metrics, the settings for the metrics used in the experiments, and the methods for correlating the metric scores with human scores. Ultimately, the benchmark will produce a table of results that shows the correlation score for each evaluated metric.

4.1 Systematic Review of Non-factoid QA

The first stage of this research is a systematic review on the topic of non-factoid QA, which was one of the earliest works in the literature to exclusively analyze studies of QA systems for non-factoid questions (CORTES et al., 2022). One of the purposes of this stage was to provide ideas and clarifications for the subsequent stages of this research. For example, among the different results presented, this review brings insights about the need for evaluating long answers automatically considering specific criteria, as some of the studies analyzed used manual evaluation methods (CAO et al., 2011; NIE et al., 2017; COSTA; KULKARNI, 2018; PITHYAACHARIYAKUL; KULKARNI, 2018), which allows for assessing specific criteria in the answers.

Regarding the metrics used to evaluate long answers, there was almost an exclu-

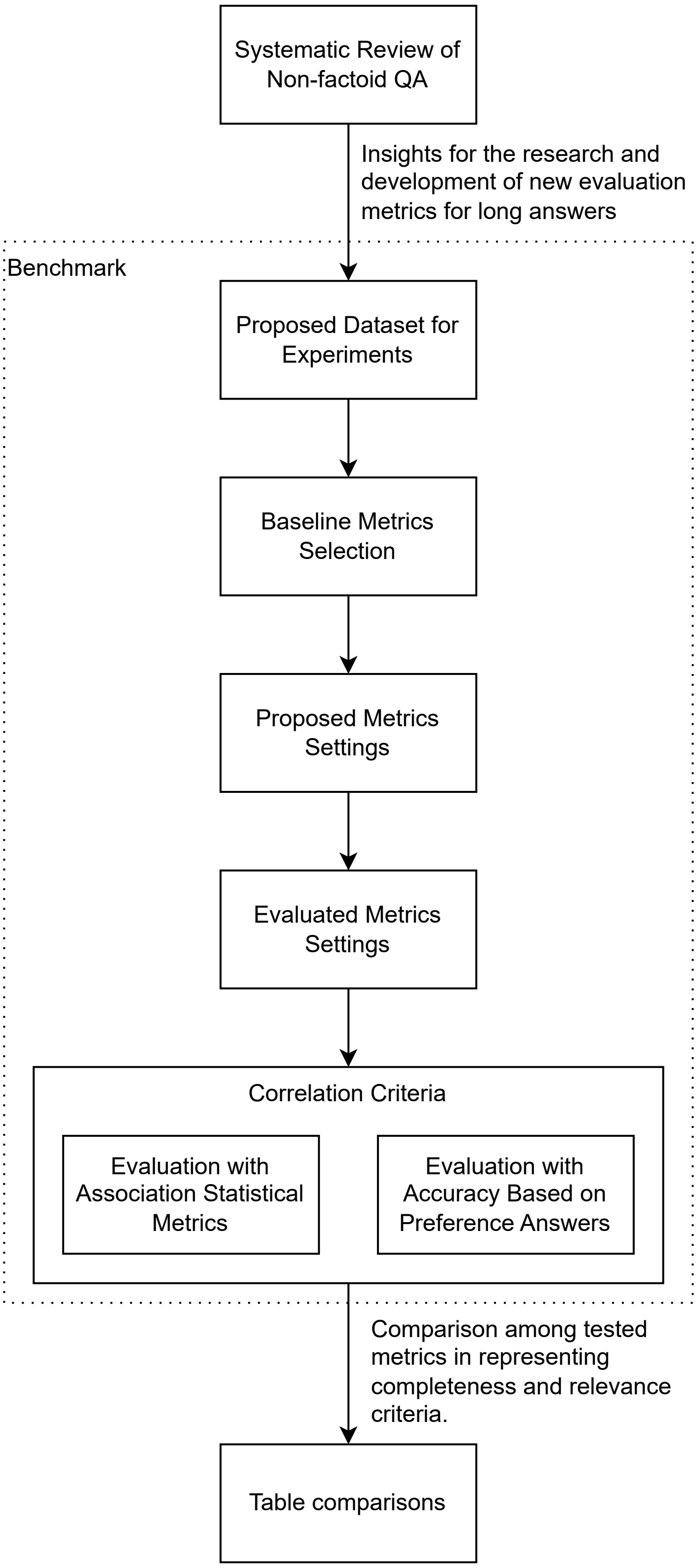

Systematic Review of Non-factoid QA
Insights for the research and development of new evaluation metrics for long answers
Benchmark
Proposed Dataset for Experiments
Baseline Metrics Selection
Proposed Metrics Settings
Evaluated Metrics Settings
Correlation Criteria
Evaluation with Association Statistical Metrics
Evaluation with Accuracy Based on Preference Answers
Comparison among tested metrics in representing completeness and relevance criteria.
Table comparisons

Figure 4.1 – Schema of the methodology.
Source: The Author.

sivity in using the ROUGE metric. It is known that ROUGE is mainly based on the exact match of words or sequences of words (n-grams) between the generated answer and a reference. This may not adequately capture the quality of longer answers, where paraphrasing, restructuring of information, and creative expression can occur. Moreover, the effectiveness of the ROUGE metric depends on the quality of the reference answers. In cases where the reference answers are not comprehensive or of high quality, the evaluation may be imprecise. The methodology of the systematic review and its results are detailed in Chapter 3.

Due to the limitations of the ROUGE metric, newer metrics such as BERTScore and BARTScore have been developed to better capture semantic similarities between the generated text and reference text. These metrics use deep learning models, such as BERT and BART, which are designed to understand the context and meaning of words in sentences, thus allowing for a more nuanced evaluation of text quality beyond just word overlap. However, these metrics does not clearly demonstrate specific evaluation criteria of the text, such as relevance, and completeness. This highlights the ongoing challenge in developing automatic evaluation metrics that can assess the specific quality aspect of long answers.

4.2 Proposed Dataset for Experiments

In this study, the dataset employed for experiments is the proposed in Chapter 5, that plays a critical role in evaluating the metrics used for analyzing the completeness and relevance of long answers. This dataset was created including only “Instruction” type questions within the domain of computer science. These questions typically begin with “How to” and demand detailed instructional content. This selection criteria allows for a focused analysis of ansers within a specific knowledge domain, reducing variability and enhancing the precision of the metrics evaluations (GERSTENBERGER et al., 2017).

The dataset comprises 106 questions derived from the “Explain Like I’m Five” (ELI5) subreddit, a part of a broader collection known for long, explanatory answers that simplify complex topics. Each question in the dataset is associated with two answers: one generated by GPT-4, and the other the highest upvoted answer from the Reddit community. This dual answer setup enables a comparative analysis of human versus AI generated content in terms of completeness and relevance.

Annotations for the dataset were conducted using a specially developed tool, allowing evaluators to score answers on a scale from 0 to 100 for both relevance and completeness. Evaluators were also tasked with identifying irrelevant portions of the content to support their relevance scores and indicating any missing information for completeness assessments. Each question in the dataset was evaluated by multiple human experts, aiming robustness in the evaluation process.

During the experiments detailed in this study, each evaluated metric is required to assign a completeness and relevance score to each answer within the dataset. These scores are then compared against the average scores provided by human annotators. This comparison enables the determination of which metrics correlate most closely with human annotations through correlation criteria. The primary aim of these experiments is to identify metrics that best mirror the scoring behavior of human annotators. Metrics that show a higher correlation with human judgment are considered more effective and are ranked accordingly. This approach aims to show the experiment's focus on aligning automated evaluation methods with human evaluative standards, ensuring that the metrics used in the assessment of answer quality reflect human perceptions of relevance and completeness.

The experiments with the dataset are aimed to set a solid foundation for future expansion to other domains or question types. More details about the creation of the dataset are presented in Chapter 5.

4.3 Baseline Metrics Selection

After creating the dataset, common metrics were selected for the evaluation of long answers that serve as a comparison basis for the metrics proposed in this research. Thus, these metrics are used to assign a score to each answer in the evaluation dataset, and then these values are correlated with the averages of the values from the human evaluators' annotations.

Although most baseline metrics generally seek to evaluate the quality of the answers and not the specific criteria of completeness and relevance, calculating the correlation with human evaluations will make it possible to analyze how much these metrics consider the criteria of relevance and completeness in their scoring. In other words, even though these metrics are known and already assessed in answer evaluation, the correlation assessment of this study could individually show how much these metrics indicate how relevant and complete an answer is.

For the selection of baseline metrics, the most commonly used metrics for evaluating the quality of long answers generated by QA models were considered. Primarily, analyses from systematic review studies (BIDGOLY; AMIRKHANI; BARADARAN, 2022; LI et al., 2023; WANG et al., 2023) were taken into account, indicating a trend towards using common metrics in translation and summarization tasks, such as BLEU and ROUGE. Also, more recent works have been using metrics based on the training of supervised models (BOLOTOVA-BARANOVA et al., 2023), like BERTScore. In addition, metrics from related works, such as RankGen (KRISHNA et al., 2022), were primarily considered. Thus, the metrics known in the literature for text evaluation and that were used in the experiments of this research are listed below:

- **BLEU**: Evaluates the quality of texts by comparing n-grams of the translated text with the n-grams of the reference text, calculating the precision of matches, adjusted by a penalty for answers shorter than their references.
- **BLEURT**: Uses a pre-trained language model to understand the context and semantics of the text. BLEURT takes into account semantic adequacy and text fluency, offering a more sophisticated evaluation than metrics based solely on n-gram overlap.
- **ROUGE**: Compares the overlap of n-grams, word sequences, and subsequences between the generated content and a set of references, focusing on the ability to capture essential information.
- **BERTScore**: Based on contextualized embeddings generated by models like BERT, BERTScore compares the semantic similarity between tokens of the generated text and the reference text, using cosine similarity between embeddings, allowing for a more refined evaluation of textual quality.
- **BARTScore**: Similar to BERTScore, but uses the BART model to evaluate the quality of generated texts, such as translations, summaries, or other text generation tasks. BARTScore can consider both the fidelity and fluency of the generated text, offering a detailed analysis of its quality.
- **RankGen**: Uses a large-scale contrastive learning approach, employing an encoder model that maps text prefixes and their continuations (model generations) to a shared vector space. The score of a generation is calculated through the inner product between the prefix vector and the generation vector.
- **CosineDistance**: A generic similarity metric that measures the angle between two vectors in multidimensional space. By measuring cosine distance, it is possible to

evaluate the semantic proximity between texts.

- **TopicDiversity**: Aimed at evaluating the diversity of topics within a set of documents or generated texts. TopicDiversity can help understand how varied the subjects addressed are, providing a view on the scope and thematic richness of the produced content.

Metrics based on the QAEval model (FABBRI et al., 2022; DEUTSCH; BEDRAX-WEISS; ROTH, 2021), which utilize question and answer generation models, were not included in the experiments due to technical limitations that resulted in inconsistent performance. The consistency and performance of this approach depend on the models that generate the question and answer pairs from the reference summary and on the model that evaluates the answers with another QA model. The models tested in this research often led to errors and incorrect answers. The inconsistency in question generation and the accuracy of the evaluated answers mean that this metric did not achieve a minimum acceptable performance quality. To ensure the reliability of the experiment results, it was decided to exclude the metric, with plans to explore improved QA models for that in future research.

In addition to the metrics known in the literature, special metrics were used that serve exclusively as a baseline for this work to analyze the influence of specific factors and show trends in the quality evaluations of the assessed answers. Using these special metrics, the study seeks to identify patterns and biases in human evaluations, as well as the relevance of certain aspects, such as the origin of the answer and its length. Thus, it is possible to gain insights into what evaluators value in the answers. These metrics are called special because they are not commonly applied in real scenarios, as they use privileged information, such as the origin of the answer and information on the size of other evaluated answers. The special metrics are described in the listing below:

- **Random**: assigns a random value from 0 to 100 to each evaluated answer. This metric is used to check how close the other metrics come to a random evaluation and also to validate the evaluation methodology of the metrics since this metric should have a low correlation with human evaluation.
- **Length**: assigns a value from 0 to 100 considering the length of the answer, where the longer the answer, the higher the score of this metric. This metric is used mainly to verify how much the length of the answer influences human evaluation. The criterion of completeness, in particular, may have a high correlation with this special

metric.

- **Always Human**: this metric always assigns a higher value to the human answer than to the answer generated by the GPT. The main idea of this metric is to verify the preference for human answers by evaluators.
- **Always GPT**: this metric always assigns a higher value to the GPT model answer than to the answer made by a human. The main idea of this metric is to verify the preference for GPT answers by evaluators.

4.4 Proposed Metrics Settings

In addition to the baseline and special metrics, the metrics proposed in this research were evaluated, which are listed below:

- **Prompt + GPT-4**: Uses the prompt strategy with in learning context to show the GPT-4 model how to evaluate answers considering the criteria of relevance and completeness. Thus, GPT-4 provides a value from 0 to 100 for each criterion.
- **Prompt + GPT-3.5**: Similarly to Prompt + GPT-4, it uses the same prompt strategy, but this time, utilizing the GPT-3.5 model.
- **Unities**: Uses the strategy of dividing the content of the answers into semantic units and calculates completeness and relevance through recall and precision.
- **Supervised**: Utilizes a supervised regression model trained with a synthetic dataset to evaluate completeness and relevance.

More details about the proposed models are presented in Chapter 6.

4.5 Evaluated Metrics Settings

The settings adopted for the evaluation of metrics aim to promote a fair and impartial evaluation process. Therefore, it is also important to consider whether the metric requires complex resources to calculate the score, and during the evaluation of the metrics, the type of resources needed to calculate the score should be considered. Thus, four classifications were used to classify each metric: i) **reference-based**, which requires a reference answer to compare with the answer predicted by the models; ii) **documents-based**, which uses documents with information that may be relevant to the construction of the answer; iii) **question-based**, which use only the text of the input question to calculate the score of the answer; iv) **answers-based**, which use only the text of the answer to be evaluated and no other resources for calculating the score of this answer. Table **??** presents the classification of the metrics used in the experiments.

Metric	**Type**
BLEU[question]	question-based
BLEU[reference]	reference-based
BLEURT[question]	question-based
BLEURT[reference]	reference-based
ROUGE[question]	question-based
ROUGE[reference]	reference-based
BERTScore[question]	question-based
BERTScore[reference]	reference-based
BARTScore[question]	question-based
BARTScore[reference]	reference-based
RankGen	question-based
CosineDistance[question]	question-based
CosineDistance[reference]	reference-based
TopicDiversity	answers-based
Prompt + GPT-4	question-based
Prompt + GPT-3.5	question-based
Regression model	question-based
Unities	documents-based

Table 4.1 – Type of metrics used in the experiments.
Source: The Author.

Although the essence of reference-based metrics requires a reference answer, the use of the text of the question itself as a reference was also tested. The idea is to check the performance of these metrics when using only the text of the question since a reference answer is often not available. Thus, if the metric shows satisfactory performance using only the text of the question, this metric could be considered for situations where there is no available reference answer. The suffix “[reference]” indicates that the metric is using a reference answer, while “[question]” indicates that the metric is using the question itself as a reference.

Considering the classifications of metrics, the reference-based class can be considered the class that requires more difficult resources to be available for an evaluation, since this type of metric requires a reference answer for each dataset instance. This reference answer must be as “correct” as possible, as during the scoring calculations, when this reference is considered entirely correct. For example, if the reference answer contains irrelevant or false sentences, answers with these same sentences will benefit. Similarly, if the reference answer does not contain all relevant information (incomplete), answers evaluated with this missed relevant information will be penalized even if they are complete. Therefore, for a fair evaluation with reference-based metrics, a set of questions with quality reference answers is necessary, which is often not possible due to the difficulty of the availability of this type of resource.

Document-based metrics can be considered easier to use compared to reference-based ones, as they do not require a specific, checked, and quality resource for each question in the evaluated dataset. Although they need documents relevant to the question, these can be extracted from existing datasets, such as the web, and automatically searched through search engines, making the process less dependent on manually verified resources specific to each question. Moreover, document-based metrics allow for a more flexible and adaptable evaluation, as the selection of relevant documents can be adjusted to better reflect the information necessary to adequately answer the question. This can include a wide range of documents, from scientific articles to blog posts and news articles, depending on the context of the question. However, this approach still depends on the quality of the search engine and the availability of relevant documents for all the questions in the evaluation dataset.

Question-based metrics are even simpler to implement, as they rely exclusively on the text of the question, eliminating the need to search for and validate external resources. Similarly, answer-based metrics represent the most independent form of evaluation, focusing only on the content of the answer without reference to question text or any external resource. This can be particularly useful in scenarios where the availability of reference data is limited or non-existent. However, scoring answers with these types of metrics is more challenging, as it relies solely on the question text or the answer text, without any external reference support, which may result in lower performance compared to metrics that consider these external resources.

4.6 Correlation Criteria

Two methods were used to determine how correlated a metric's ability is in assigning relevance and completeness scores in correlation with the scores of human annotators. The first uses correlation coefficients to measure the strength and direction of the association between two variables. The second calculates the accuracy of the metrics by verifying if the answer with the highest score assigned by the metric was the same answer chosen with the highest score assigned by human evaluators.

In both forms of evaluation, the average scores among annotators for the criteria of relevance and completeness were considered. Thus, for each evaluated answer, the arithmetic mean of the scores from the four evaluators for each criterion was performed. The resulting distribution can be seen in Figure 5.5 for the relevance criterion, and in Figure 5.6 for the completeness criterion, in Chapter 5.

4.6.1 Evaluation with Association Statistical Metrics

The evaluation of the proposed and baseline metrics was conducted using statistical association metrics to quantify the correlation between the scores assigned by the metrics and the evaluations of human annotators. For this purpose, three correlation coefficients were used: Spearman's rank, Kendall's rank, and Pearson. To do this, two aligned lists of scores for each criterion (relevance and completeness) for each answer in the test dataset are created, one with the scores from the evaluated metric and the other with the scores from human evaluators. From these two lists, the functions of the correlation coefficients are applied and return the statistical value. Each of these statistical metrics offers a unique perspective on the association between the datasets, taking into account different aspects of the correlation.

For each of these correlation coefficients, the returned value ranges from -1 to 1, where:

- **1** indicates a perfect positive correlation, meaning that as one variable increases, the other also increases.
- **0** (zero) indicates that there is no correlation between the variables.
- **-1** indicates a perfect negative correlation, which implies that as one variable increases, the other decreases.

In addition to the correlation value, a p-value is also obtained, which indicates the probability of observing the calculated correlation in a world where the true correlation is zero. A p-value lower than a threshold (usually, 0.05) indicate that the observed correlation is statistically significant, suggesting that there is a real association between the variables, and not a result that occurred by chance.

4.6.1.1 Spearman

The Spearman correlation, also denoted as r_s, is a non-parametric statistical measure that evaluates the dependence or association between the rankings of two variables. This coefficient is used to analyze how well the relationship between two variables can be described by a monotonic function. Unlike Pearson's correlation, which assesses the linear relationship between variables, Spearman considers the order of the values and is more suitable for non-parametric data. It is particularly useful when the data do not follow a normal distribution or when the relationship between the variables is not linear.

To calculate the Spearman coefficient, the raw values of the two variables are sorted in ascending order and a rank is assigned to each value, from the lowest to the highest. For each pair of values, the difference in their ranks is calculated. Then, the formula for Pearson's correlation is used, but applied to the ranks instead of the original values. Thus, two variables X_i and Y_i are converted to ranks $R(X_i)$ and $R(Y_i)$, and the Spearman coefficient is calculated using the equation below:

$$r_s = \rho_{R(X),R(Y)} = \frac{\operatorname{cov}(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

where r_s is the symbol for the Spearman correlation coefficient, ρ is the usual Pearson correlation coefficient but applied to the ranks of the variables. $\operatorname{cov}(R(X), R(Y))$ is the covariance of the ranks of the variables, and $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the standard deviations of the ranks of the variables.

The simplified formula for cases where all ranks are unique is shown in the equation below:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks for each observation, and n is the number of observations.

For experiments evaluating metrics, the Spearman coefficient is one of the main indicators of correlation between the metrics and human evaluation, as it deals with rank values, which is important for evaluation, since different metrics may use different scales of values and may not be normalized.

4.6.1.2 Kendall

The Kendall correlation, also denoted as τ, is a non-parametric measure that evaluates the strength of the dependency between two variables. This coefficient is based on the agreement between pairs of observations. Like Spearman's correlation, Kendalltau is suitable for non-parametric data and is useful when the relationship between the variables is monotonic, but not necessarily linear. Its advantage lies in the intuitive interpretation, even in small datasets.

The Kendall coefficient assesses the relationship between two variables by comparing pairs of observations. A pair of observations is considered concordant if the relative order of both observations is the same in both variables. For example, if in two lists, an item is above another in one list and also above in the other list, that pair is concordant. In the other hand, a pair is discordant if the relative order in the two variables is inverse. The Kendall coefficient τ is calculated by the difference between the number of concordant and discordant pairs, divided by the total number of possible pairs, as shown in the equation below:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}$$

where $\binom{n}{2}$ represents the total number of possible pairs to be chosen from n items, which is calculated as $n(n-1)/2$.

In experiments evaluating metrics, the Kendall coefficient also deals with rank values and is used alongside the Spearman coefficient to determine the correlation between the scores of the evaluated metrics and the human score. Thus, by using both coefficients, it is possible to gain a more comprehensive understanding of the relationship between the metric scores and human scores.

4.6.1.3 Pearson

The Pearson correlation is a statistical measure that assesses the degree of linear correlation between two quantitative variables. In other words, it quantifies the strength and direction of the linear relationship between these variables.

The formula for calculating the Pearson correlation coefficient between two variables X and Y is given by the ratio of the covariance of X and Y to the product of the standard deviations of X and Y. This normalizes the covariance measure, allowing for a direct interpretation of the strength of the linear relationship. Its equation is presented below and is indicated when applied to a population:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\text{cov}(X,Y)$ is the covariance between the variables X and Y, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y.

Although widely used, Pearson recommend that the data follow a normal distribution and that the relationship between the variables is linear, which may not be the case in all scenarios. Therefore, in the case of this study, it is important to complement its application with Spearman and Kendall measures, which are less restrictive regarding the nature of the data.

4.6.1.4 Answers Subset Evaluations

In addition to the evaluation using all the answers from the evaluation set, the answers generated by humans (Reddit users) and those generated by the GPT-4 model are also evaluated individually. Thus, three evaluation contexts utilize the three correlation coefficients: the first involves the evaluation of all 212 answers generated by both humans and GPT-4; the second uses only the 106 answers created by humans; and the third uses only the 106 answers created by the GPT-4 model.

Answers generated by different sources, such as those from humans and the GPT-4 model, may present various characteristics that differentiate them, such as those related to writing style and the sources used. The answers from human Reddit users can be more informal, containing specific vocabulary from the forum, and content like links to videos and images about the topic. Also, these answers may exhibit a strong position about a subject. On the other hand, answers generated by the GPT-4 model are typically formal, with clear formatting and seeking to be objective. Also, the GPT-4 model's answers may contain a more neutral position.

Given that the answers generated by different sources contain characteristics that differentiate them, the individual evaluation of the metrics can show significant results when analyzing their performance in the three different evaluation contexts (all answers, human answers, and GPT-4 answers). Furthermore, this approach allowed for a detailed analysis of the metrics' performance in different scenarios, highlighting their capabilities in specific contexts and identifying potential biases or trends in human evaluations.

4.6.2 Evaluation with Accuracy Based on Preference Answers

This subsection present the methodology employed to assess the accuracy of the evaluated metrics by comparing their preference for answers against the preferences indicated by human annotators. This approach goes beyond correlating scores; it directly investigates whether the metrics and human annotators agree on which of the two answers (Reddit user or GPT-4 generated) is more relevant or complete for each question.

To calculate accuracy, we compare the preference of each metric with the preference shown by human evaluators. For each question, we have two answers: one generated by a human (Reddit user) and the other by the GPT-4 model. Both the evaluated metrics and human annotators assign scores for relevance and completeness to each answer. The accuracy of a metric is then determined by the percentage of instances where the metric's preferred answer (the one with the higher score) matches the human annotators' preferred answer for the same question. This comparison is done separately for relevance and completeness criteria.

The preference for an answer by both metrics and human annotators is determined based on which answer receives the higher score. If the score assigned by a metric or human annotators to the Reddit user's answer is higher than that assigned to the GPT-4's answer, the Reddit user's answer is considered preferred, and vice versa.

To statistically validate the accuracy results, we employ the binomial test. This choice is justified as the accuracy measurement involves binary outcomes (correct or incorrect preference match) for each question. The binomial test assesses whether the observed proportion of correct matches significantly deviates from the chance level (50$\%$, assuming an equal likelihood of preferring either answer).

The hypothesis tested in this context is: - H_0: The probability of a metric's preference matching the human annotators' preference is equal to 0.5 (chance level). - H_a: The probability of a metric's preference matching the human annotators' preference is different from 0.5.

A p-value is calculated for each metric, indicating the likelihood of observing the obtained accuracy under the null hypothesis. A low p-value (typically $p < 0.05$) suggests that the metric's accuracy in matching human preference is significantly better than chance, providing evidence of the metric's effectiveness in assessing answer quality according to human judgments.

This method provides a simple way to determine how well different metrics match human opinion. Instead of comparing numerical scores, it looks at which answer annotators prefer. By focusing on which answers are preferred, it is possible to see which metrics are in tune with human evaluation.

5 COMPLETENESS AND RELEVANCE ANNOTATED ANSWERS DATASET

This chapter introduce a specialized dataset focused on “Instruction” type questions in the field of computer science composed by long answers annotated by humans for the completeness and relevance criterias. This dataset should serves as a foundational tool for evaluating metric models focused in completeness and relevance criterias, aiming a controlled environment to reduce biases and enhance interpretability.

For the evaluation of the metrics, a dataset was created containing questions, answers, and human annotations for completeness and relevance criteria. A decision was made to use only one class of question and one knowledge domain. In this case, “Instruction” type questions were chosen, which typically start with “How to...” and require steps with instructions and explanations. The “Instruction” type is a class of question that is easy to identify and common in datasets with long answers. As for the knowledge domain, the field of computer science was chosen, as it is common to find instructional type questions in this domain on public datasets. These observations were made previously in empirical analysis of datasets with long answers.

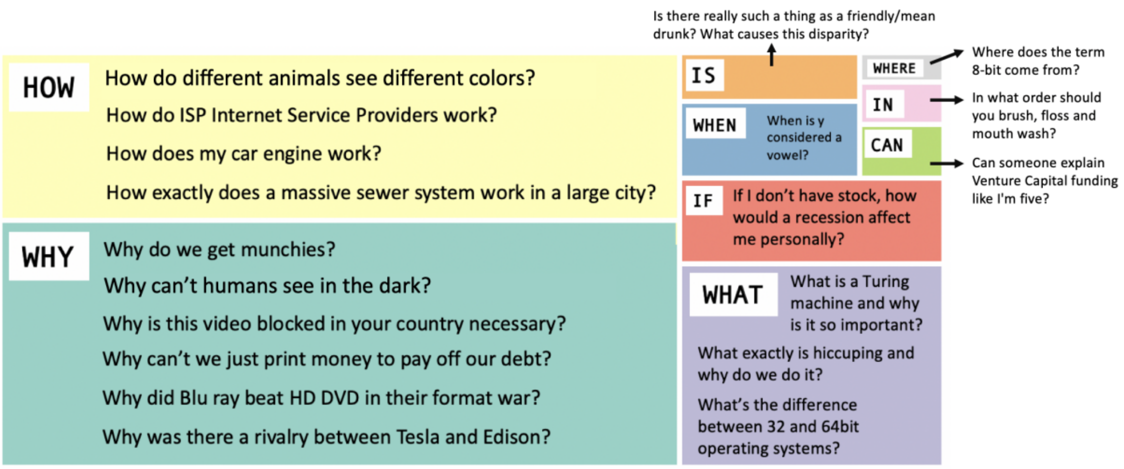
The choice to specify the type of question and the knowledge domain was made mainly to allow the focus of annotation resources on a single domain, enabling the use of different human experts to evaluate the same answers, thus increasing the reliability of the annotations. Therefore, the metrics evaluated and developed are initially validated in a specific domain for instructional type questions, where the answers and annotations are more controlled and standardized. One of the advantages of this approach is to provide a solid basis for testing with specific data, aiming to reduce biases and common errors. Moreover, it also allows for easier analysis and interpretation, as the dataset is more homogeneous. Finally, starting with a specific domain establishes a clearer and more solid starting point, and from there, gradually expand and adapt the metric to other types of questions and domains, adjusting it as necessary.

5.1 Question and answers extraction from ELI5

The first step in creating the dataset was to select questions from the ELI5 (Explain Like I’m Five) collection, which consists of questions that require long, explanatory answers, with the goal of facilitating the understanding of complex topics. The choice of this collection is primarily due to the characteristic of long answers, as shown in Table **??**,

Dataset	# of Words in Question	# of Words in Answer	# Q-A Pairs
ELI5	42.2	130.6	272K
MS MARCO v2	6.4	13.8	183K
TriviaQA	14	2.0	110K
NarrativeQA	9.8	4.7	47K
CoQA	5.5	2.7	127K
SQuAD (2.0)	9.9	3.2	150K
HotpotQA	17.8	2.2	113K

Table 5.1 – Comparing large-scale QA datasets about question and answer words magnitudes. ELI5 have longer questions and answers.
Source: adapted from (FAN et al., 2019)



How do different animals see different colors?
How do ISP Internet Service Providers work?
How does my car engine work?
How exactly does a massive sewer system work in a large city?

Why do we get munchies?
Why can't humans see in the dark?
Why is this video blocked in your country necessary?
Why can't we just print money to pay off our debt?
Why did Blu ray beat HD DVD in their format war?
Why was there a rivalry between Tesla and Edison?

Is there really such a thing as a friendly/mean drunk? What causes this disparity?

When is y considered a vowel?

Where does the term 8-bit come from?

In what order should you brush, floss and mouth wash?

Can someone explain Venture Capital funding like I'm five?

If I don't have stock, how would a recession affect me personally?

What is a Turing machine and why is it so important?
What exactly is hiccuping and why do we do it?
What's the difference between 32 and 64bit operating systems?

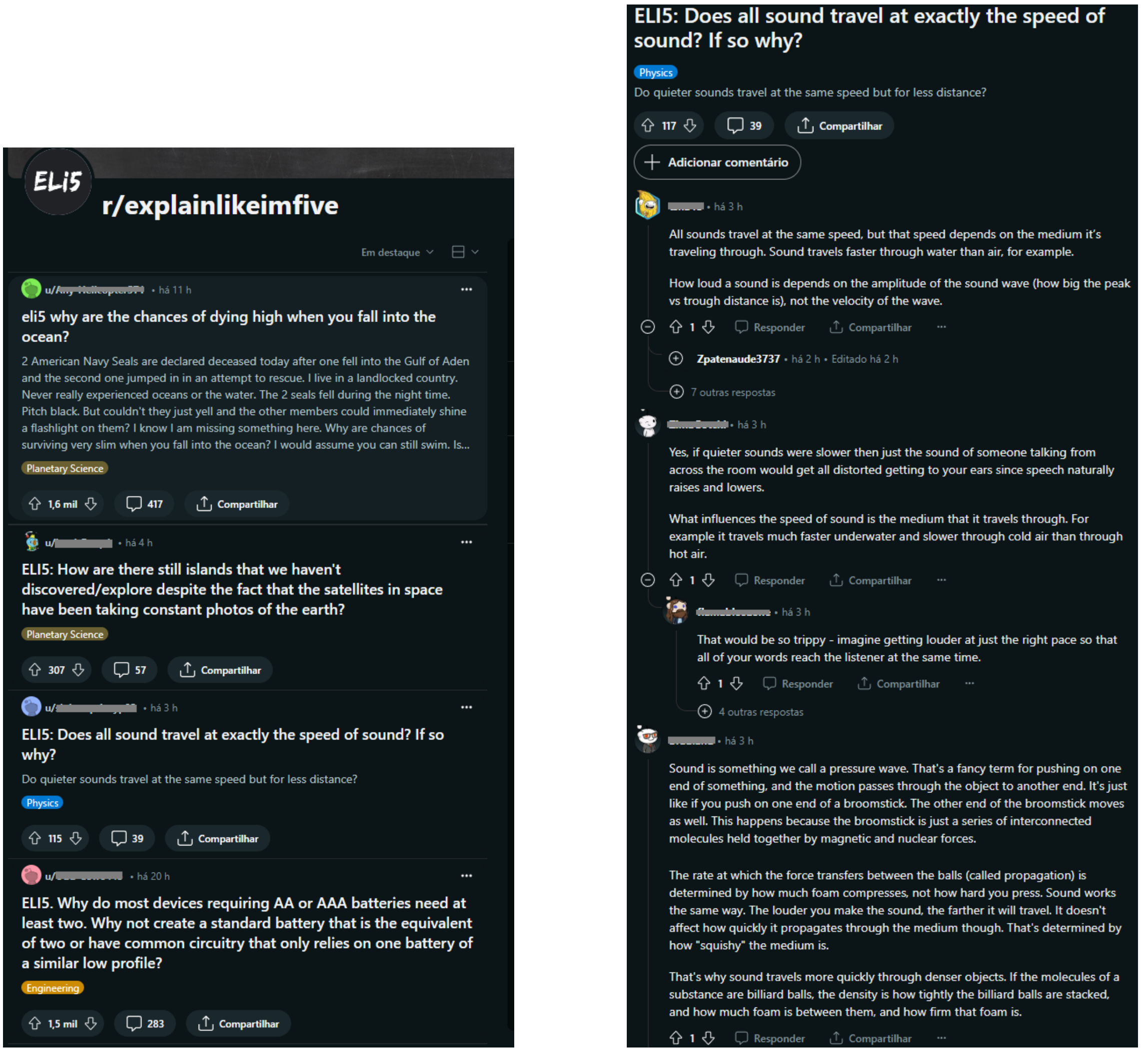
Figure 5.1 – Question first words exanples from ELI5, where box size represents frequency.
Source: from (FAN et al., 2019)

where ELI5 presents considerably longer answers than other known datasets. Additionally, this collection has a large volume of questions that can be classified as "Instruction", as shown in Figure 5.1, where questions starting with the word "HOW" can be classified as such.

These questions are found on a subreddit, as shown in Figure 5.2. It is a platform where users post questions about various topics, seeking simplified explanations, as if they were explaining to a five-year-old child. User interaction on the ELI5 subreddit is dynamic and collaborative, with community members answering to questions with detailed and accessible information. The answers are then voted on by community users, with the most voted ones being highlighted.

The questions and answers from the ELI5 subreddit form the basis of this dataset and were extracted using a script from the ELI5 project[1] (FAN et al., 2019). This is a detailed process that involves several steps of data processing and filtering. The need to run scripts instead of directly downloading the data is due to licensing restrictions and the

[1] www.github.com/facebookresearch/ELI5



u/[redacted] • há 11 h

eli5 why are the chances of dying high when you fall into the ocean?

2 American Navy Seals are declared deceased today after one fell into the Gulf of Aden and the second one jumped in in an attempt to rescue. I live in a landlocked country. Never really experienced oceans or the water. The 2 seals fell during the night time. Pitch black. But couldn't they just yell and the other members could immediately shine a flashlight on them? I know I am missing something here. Why are chances of surviving very slim when you fall into the ocean? I would assume you can still swim. Is...

Planetary Science

u/[redacted] • há 4 h

ELI5: How are there still islands that we haven't discovered/explore despite the fact that the satellites in space have been taking constant photos of the earth?

Planetary Science

u/[redacted] • há 3 h

ELI5: Does all sound travel at exactly the speed of sound? If so why?

Do quieter sounds travel at the same speed but for less distance?

Physics

u/[redacted] • há 20 h

ELI5. Why do most devices requiring AA or AAA batteries need at least two. Why not create a standard battery that is the equivalent of two or have common circuitry that only relies on one battery of a similar low profile?

Engineering

ELI5: Does all sound travel at exactly the speed of sound? If so why?

Physics

Do quieter sounds travel at the same speed but for less distance?

[redacted] • há 3 h

All sounds travel at the same speed, but that speed depends on the medium it’s traveling through. Sound travels faster through water than air, for example.

How loud a sound is depends on the amplitude of the sound wave (how big the peak vs trough distance is), not the velocity of the wave.

Zpatenaude3737 • há 2 h • Editado há 2 h

7 outras respostas

[redacted] • há 3 h

Yes, if quieter sounds were slower then just the sound of someone talking from across the room would get all distorted getting to your ears since speech naturally raises and lowers.

What influences the speed of sound is the medium that it travels through. For example it travels much faster underwater and slower through cold air than through hot air.

[redacted] • há 3 h

That would be so trippy - imagine getting louder at just the right pace so that all of your words reach the listener at the same time.

4 outras respostas

[redacted] • há 3 h

Sound is something we call a pressure wave. That's a fancy term for pushing on one end of something, and the motion passes through the object to another end. It's just like if you push on one end of a broomstick. The other end of the broomstick moves as well. This happens because the broomstick is just a series of interconnected molecules held together by magnetic and nuclear forces.

The rate at which the force transfers between the balls (called propagation) is determined by how much foam compresses, not how hard you press. Sound works the same way. The louder you make the sound, the farther it will travel. It doesn't affect how quickly it propagates through the medium though. That's determined by how "squishy" the medium is.

That's why sound travels more quickly through denser objects. If the molecules of a substance are billiard balls, the density is how tightly the billiard balls are stacked, and how much foam is between them, and how firm that foam is.

(a) Reddit page with questions. (b) Reddit page with answers for a question.

Figure 5.2 – ELI5 subreddit.

Source: www.reddit.com/r/explainlikeimfive complexity of the processing required. For the first step in creating the dataset, questions and answers from the period of July 2011 to July 2018 were extracted. In this dataset, one answer is attributed to each question, which in this case are the answers with the highest user vote scores.

5.1.1 Filtering by question type

The original dataset from ELI5 extracted contains 307,018 instances of questions and answers. To determine which of these questions are of the "Instruction" type, the model *Lurunchik/nf-cats*[2] trained with the NF-CATS dataset (BOLOTOVA et al., 2022), containing a taxonomy of non-factoid questions, was used. This model allows for the identification of questions from the "Instructio" class with a performance of 94.3% for the F1-score metric. Through this model, 2,408 questions classified as "Instruction" were extracted.

5.1.2 Filtering by domain

After extracting questions of the "Instruction" type, questions from the field of computer science were selected by ranking them based on their similarity to the topic. To determine this similarity, the cosine similarity method was used, where the embeddings of the questions were compared with the embeddings of terms related to computer science. The embeddings were represented using the *bert-base-uncased* model (DEVLIN et al., 2018). After ranking, 106 questions were manually selected. This selection only considered questions with answers that were not too short (more than 10 words) and not too long (less than 500 words). These criteria are related to the process of annotation by human evaluators, where an adequate number of questions were selected considering the available resources for annotation with different evaluators. Finally, minor modifications were made to the questions to make them clearer, involving appropriate capitalization of characters and removal of specific terms from the Reddit portal, such as "[ELI5]" present in the text of some questions.

[2] www.huggingface.co/Lurunchik/nf-cats

5.2 Reference answer creation

In the context of the metrics evaluated in this research, many are of the reference-based type and depend on a reference answer for evaluation. When considering the answers from ELI5, these are generated by the Reddit community, and their quality is automatically determined by the extraction script based on the number of user votes. However, this method has significant limitations. Firstly, the diversity of Reddit's voting users, many of whom are not experts in the topic discussed, can compromise the accuracy of the most upvoted answers. Additionally, attractive or well-written answers, including those with humorous content, often receive more votes, which does not necessarily reflect greater relevance or completeness. There's also the tendency for users to vote for answers that confirm their own beliefs or understandings, regardless of their truthfulness. Finally, the bandwagon effect is another concern: answers with many initial votes can attract additional votes simply because of their apparent popularity, especially among less experienced users.

Given the problems with Reddit's answers, we chose to use this answer as one of the answers evaluated by the evaluators in the annotation of relevance and completeness of the answer. Therefore, we opted to create the reference answers through human specialists in the field of computing. For this purpose, two experts in computing were hired on the UpWork platform to create answers for the 106 selected questions. Each of the experts created their own answer to all questions, following the instructions of:

- The answers should be as relevant and complete as possible. In this case, it was explained didactically, in detail, and with examples of what the criteria for relevance and completeness would be.
- Use reliable information sources to create the answers. The link to these information sources should also be provided.
- Include only textual information and not information in other formats, such as tables and images.
- Not to rely on answers generated by LLMs, like ChatGPT. The answers from LLMs should be verified with reliable sources.
- The option was given for the user not to answer the question if they did not feel comfortable with the subject.

After the answers were created by the two experts, a third expert checked the two answers for each question, following the criteria of relevance and completeness, and selected the most suitable one.

5.3 Answer annotation

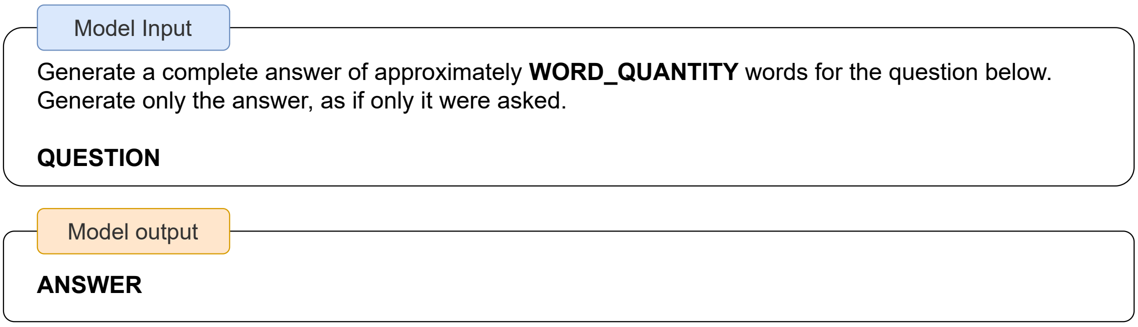
The evaluators evaluated two answers per question using a specially developed annotation tool. The process involved assessing two versions of answers for each question. For each answer, the evaluator assigned a value from 0 to 100 for the criteria of completeness and relevance. In addition to assigning the value, the evaluator had to select in the answer text which parts they considered irrelevant. This selection aims to ensure that the users are paying attention to the text in relation to what they consider relevant and irrelevant. Thus, when assigning the relevance score, they can consider the proportions of the text they consider relevant. Moreover, when the user assigns a low score for the completeness criterion, they must justify what kind of information is missing in the answer by selecting from a set of provided justifications.

5.3.1 Selecting answers for annotation

Two answers per question were used for annotation with human evaluation of the criteria for completeness and relevance. One of the answers is extracted from ELI5, representing the most upvoted answer by the community. The second evaluated answer was generated by the LLM GPT-4. Thus, for each question, we have the best-ranked human answer according to the votes of the Reddit community, and an answer generated by an LLM with results compatible with the state of the art in QA (MAO et al., 2023).

The length of the answer is an important factor when considering mainly the criterion of completeness. More complete answers are usually longer. Also, the criterion of relevance relates in such a way that longer answers are more likely to contain irrelevant information. Therefore, the answers generated by GPT-4 aim to contain the same word count as the answer created by the human from Reddit. For this, the prompt shown in Figure 5.3 was used, which specifies how many words the answer (“ANSWER”) should contain through the value specified in “WORD_QUANTITY”. The GPT-4 model was used via the API, with the temperature parameter fixed at 0.

Thus, each human evaluator assessed two answers for each of the questions, to-



Generate a complete answer of approximately **WORD_QUANTITY** words for the question below. Generate only the answer, as if only it were asked.

QUESTION

ANSWER

Figure 5.3 – Prompt for answer creation looking for answers with controlled size. WORD_QUANTITY is replaced by the words quantity and QUESTION by the input question. The model's output is the ANSWER. Source: the authors.
Source: The Author.

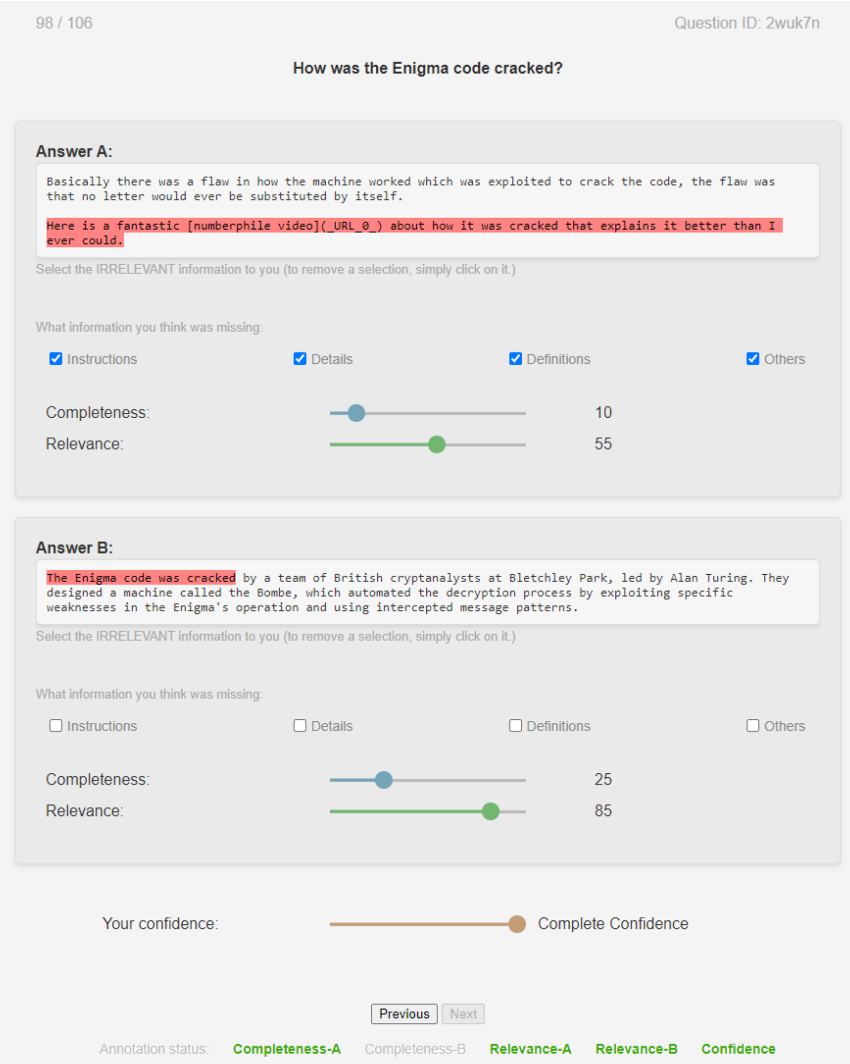
taling 212 answers evaluated for the criteria of relevance and 212 answers evaluated for completeness.

5.3.2 Annotation Answers Tool

The annotations were made using a web-based tool developed with an interface that allows the annotator to assign scores for completeness and relevance to the answers, as shown in Figure 5.4. At the top of the tool's page, the index of the current answer is displayed, enabling the user to see how many answers are yet to be evaluated. Directly below, the question for the answers is shown. The two questions, referred to as "Answer A" and "Answer B", are displayed right underneath. The order of the human answer and the GPT-4 answer is randomly assigned to each position to reduce biases related to the order.

Below each answer, a slider is provided for the user to quickly and easily indicate the value assigned for the criteria of completeness and relevance. Additionally, when the user selects a part of the answer text, it is highlighted with a red background, signifying that the segment is irrelevant. Also, the options for justifying missing information are displayed in checkboxes. At the end of the page, the user is also required to indicate their confidence level regarding the evaluation of the two answers. The button to proceed to the next question is enabled only after the user has made all the annotations. The status of the annotation can be seen in the bottom bar. In the case of the example in Figure 5.4, the user still needs to select the options that justify the low score for the completeness criterion.

The checkboxes below the label "What information you think was missing:" are



98 / 106

Question ID: 2wuk7n

How was the Enigma code cracked?

Answer A:

Basically there was a flaw in how the machine worked which was exploited to crack the code, the flaw was that no letter would ever be substituted by itself.

Here is a fantastic [numberphile video](_URL_0_) about how it was cracked that explains it better than I ever could.

Select the IRRELEVANT information to you (to remove a selection, simply click on it.)

What information you think was missing:

☑ Instructions ☑ Details ☑ Definitions ☑ Others

Completeness: 10

Relevance: 55

Answer B:

The Enigma code was cracked by a team of British cryptanalysts at Bletchley Park, led by Alan Turing. They designed a machine called the Bombe, which automated the decryption process by exploiting specific weaknesses in the Enigma's operation and using intercepted message patterns.

Select the IRRELEVANT information to you (to remove a selection, simply click on it.)

What information you think was missing:

☐ Instructions ☐ Details ☐ Definitions ☐ Others

Completeness: 25

Relevance: 85

Your confidence: Complete Confidence

Annotation status: Completeness-A Completeness-B Relevance-A Relevance-B Confidence

Figure 5.4 – Annotation tool interface. Source: the authors.
Source: The Author.

used for the user to justify the reason for a completeness score below 90. This type of annotation, along with the selection of parts of the answer that are irrelevant, serve as mechanisms that help ensure the user is paying attention to the annotations.

5.3.3 Completeness and Relevance Annotation

The annotation of answers in terms of their completeness and relevance was a pivotal part of this study, focused on assessing the long answers considering these two criteria. The dynamics of annotation were designed to provide a comprehensive evaluation of the answers, designed to ensure that each answer was correctly evaluated and annotated by the human annotators participating in the process.

Initially, the candidate annotators were subjected to a preliminary test, comprising five questions with two answers each. This test served as a test to evaluate their understanding of the completeness and relevance criteria, after read the annotation guideline. The answers evaluated by annotators in this test were built with distinct elements that clearly indicated either high or low completeness and relevance. For instance, some answers contained a significant amount of text unrelated to the question, directly impacting their relevance score. Others have incomplete information, thus affecting their completeness score. Additionally, the test served to assess the annotators' basic computer science knowledge, a prerequisite for their participation in this study. In total, 30 candidates participated in the evaluation process.

The chosen annotators, four in total, were identified based on their high correlation with reference annotations which are the ones that were as objective and less subjective as possible. These reference annotations were crucial in ensuring a standardized approach to the evaluation process. Besides their performance in the preliminary test, other factors such as their reputation on the UpWork platform and experience with similar tasks were also considered in their selection. Each annotator was compensated with $20 for evaluating the answers to 106 questions, distributed in four stages with a payment of $5 per stage, corresponding to the annotation of 26 questions.

To maintain the quality of the annotations and avoid fatigue, were given a wide window of time to annotators and them was advised not to engage in the task for more than one consecutive hour. The annotation process was structured: for each question, the annotators were presented with two answers, one from ELI5 subreddit and another generated by GPT-4. They were required to assign a score from 0 to 100 for both the

Criterion	ICC3k (Average fixed evaluators)	95% Confidence Interval
Relevance	0.5815	[0.48, 0.67]
Completeness	0.7817	[0.73, 0.83]

Table 5.2 – Intraclass Correlation Coefficient Results for Relevance and Completeness
Source: The Author.

completeness and relevance criteria for each answer.

Additionally, the annotators used the annotation guide to support their evaluation. This guide provided them with detailed explanations and examples of how to score answers based on their completeness and relevance. It also included instructions on how to identify and highlight irrelevant text segments within an answer, which was a crucial part of substantiating the relevance scores. In cases where an answer received a completeness score below 90, the annotators were required to justify their assessment by selecting from predefined options that indicated missing information types.

This structured approach to annotation aimed to provide a detailed evaluation of each answer, ensuring that the scores assigned were reflective of the annotator's thoughtful consideration of each answer's content. By examining each answer for elements of completeness and relevance, and providing justifications for their scoring, the annotators contributed to an important dataset that offers insights into the quality of answers generated by GPT-4 and humans for instructions questions about computer science considering completeness and relevance criterias.

5.4 Dataset analysis

5.4.1 Annotation correlation

To assess the consistency of evaluations conducted by four evaluators, the Intraclass Correlation Coefficient (ICC) was used. The ICC is a statistical measure that quantifies the degree of agreement among evaluators for quantitative estimates. Specifically, the ICC3k, which is suitable for situations where there is a fixed set of evaluators evaluating all instances, was employed in this case. This ICC model assumes that the evaluators are the only ones of interest and do not represent a sample from a larger group.

The ICC results for the criteria of relevance and completeness of the answers are presented in Table 5.2.

For the Relevance criterion, an ICC3k of 0.5815 was observed, reflecting moderate agreement among the raters. This level of agreement indicates reasonable harmony among the raters, especially considering the intrinsically subjective nature of relevance assessment. The relevance of an answer can vary significantly based on individual experiences and each rater's understanding of the topic. Therefore, some variation in perceptions of relevance is expected. The 95% confidence interval, ranging from 0.48 to 0.67, supports this interpretation, suggesting that while there is some variation in assessments, there is a baseline of common agreement among the raters.

In the case of the Completeness criterion, an ICC3k of 0.7817 was observed, indicating substantial agreement among the raters. This result is particularly noteworthy given the inherent subjectivity in judging the completeness of an answer. Completeness can be influenced by the raters' familiarity with the topic and their perception of what aspects are essential for a complete answer. The 95% confidence interval for Completeness, ranging from 0.73 to 0.83, shows consistency in the assessments, suggesting that despite individual differences, raters share a common understanding of what constitutes a complete answer.

In both criteria, the statistical significance of the results (p-values less than 10^{-16}) reinforces the reliability of these measurements. The 95% confidence intervals provide a range of variation within which the true ICC of the population can be expected to be contained despite the inherent subjectivity of the evaluated criteria.

The observed difference in the ICC3k values for the Relevance and Completeness criteria reflects the characteristic subjectivity associated with each of these evaluation aspects. The lower ICC3k for Relevance (0.5815) suggests moderate agreement among the raters, which can be attributed to the subjective nature of determining what is relevant in an answer. Relevance can often be influenced by the individual perceptions and experiences, leading to significant variations in their assessments. On the other hand, the Completeness criterion, with a higher ICC3k of 0.7817, indicates substantial agreement. This may be due to the fact that completeness is generally more tangible and measurable, based on the presence or absence of specific elements in the answer, thus reducing variation in raters' perceptions. For example, this facilitated perception could be related to the length of the answer, where users might associate longer answers with more complete answers.

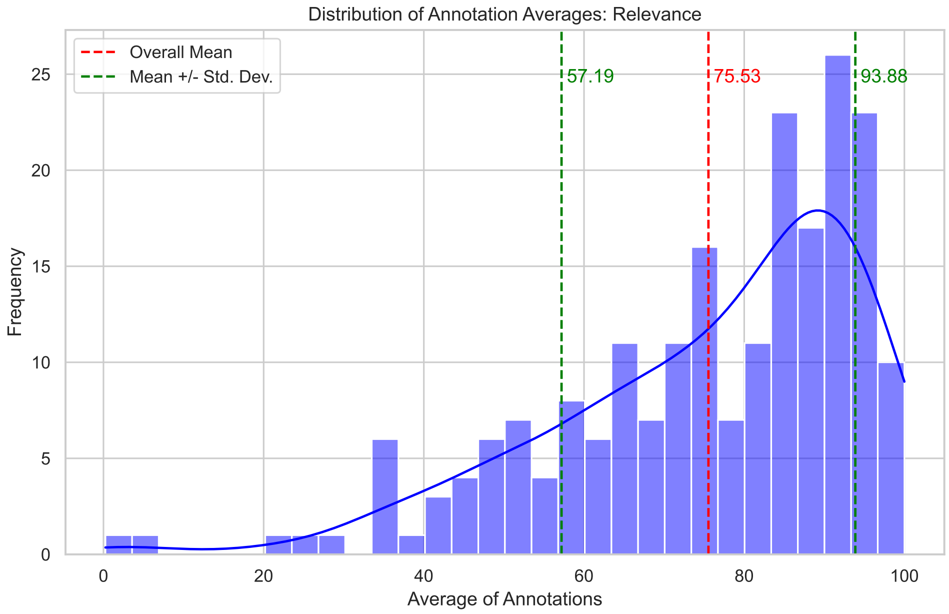


Figure 5.5 – Mean Distributions of Relevance Scores.
Source: The Author.

5.4.2 Distributions of Annotations Scores

The distribution of annotation scores is an important aspect of understanding the consistency and tendencies in the evaluations made by the annotators. Analyzing the distribution helps identifying any biases and provides insights into the overall quality of the answers as perceived by the annotators. The distributions for the criteria of Relevance and Completeness are visualized through the histograms in Figures 5.5 and 5.6, respectively, providing a graphical representation of the scores assigned by the annotators.

For the Relevance criterion, the histogram shows a distribution that is somewhat left-skewed, suggesting a concentration of scores above the mean. The overall mean score for relevance is 75.53, with a standard deviation of 18.34, indicating that on average, answers are considered relatively relevant with some variability. The presence of high scores clustered together suggests that the reviewers generally found the answers relevant to the questions. This can be partly attributed to the annotators' ability to visually identify and highlight irrelevant information within the answers, which provides a clear basis for assigning relevance scores.

In contrast, the Completeness scores' distribution appears more asymmetric with multiple peaks and also shows a left-skew. The overall mean score for completeness is 66.52, with a standard deviation of 17.28. The more varied distribution and presence of

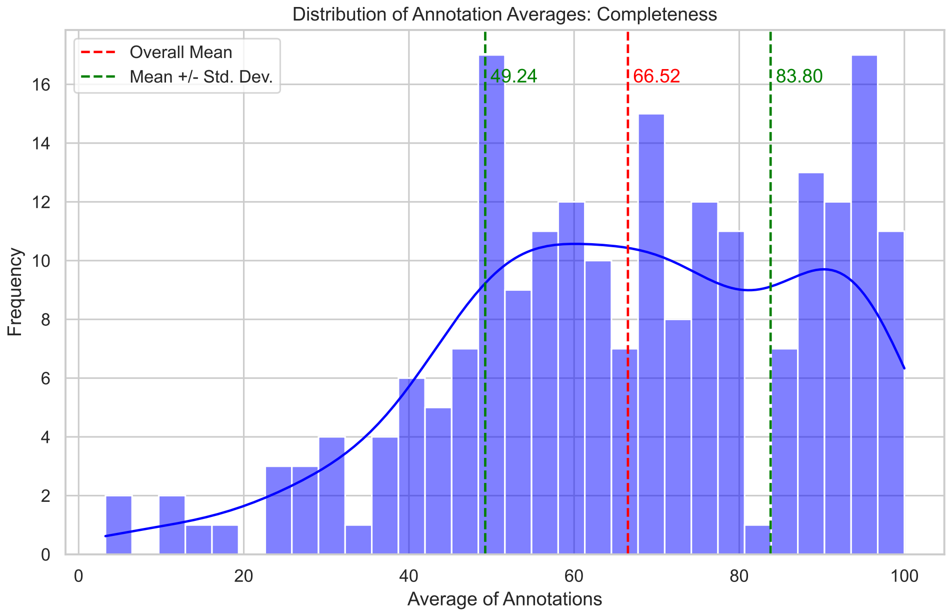


Figure 5.6 – Mean Distributions of Completeness Scores.
Source: The Author.

distinct peaks, particularly the noticeable peak at a score of 50, may indicate a tendency among annotators to choose mid-scale values when uncertain. This pattern could suggest that when annotators are unsure about the extent of completeness, they might default to a neutral score, reflecting a balance between what is present in the answer and what they perceive to be missing.

Comparing the distributions of Relevance and Completeness, it's observed that relevance scores tend to form a more normalized distribution. This is probably because it's easier to see how relevant something is when highlighting text in the annotation tool. The visual help in identifying irrelevant segments may lead to a more uniform assessment of relevance. Completeness, on the other hand, seems to be a more complex measure with a less normalized distribution compared to relevance. Unlike relevance, which can be more directly assessed through visual cues in the annotation tool, completeness evaluations are more dependent on the annotators' judgment and understanding of the topic. Because each annotators use their own judgment, there may be a wider range of scores, showing the different viewpoints each person has.

Furthermore, the higher scores for both criteria suggest that the quality of the evaluated answers was generally high. This is consistent with the nature of the dataset, where human answers were those most upvoted by the community, and the automated answers were generated by GPT-4, known for its advanced capabilities in understanding,

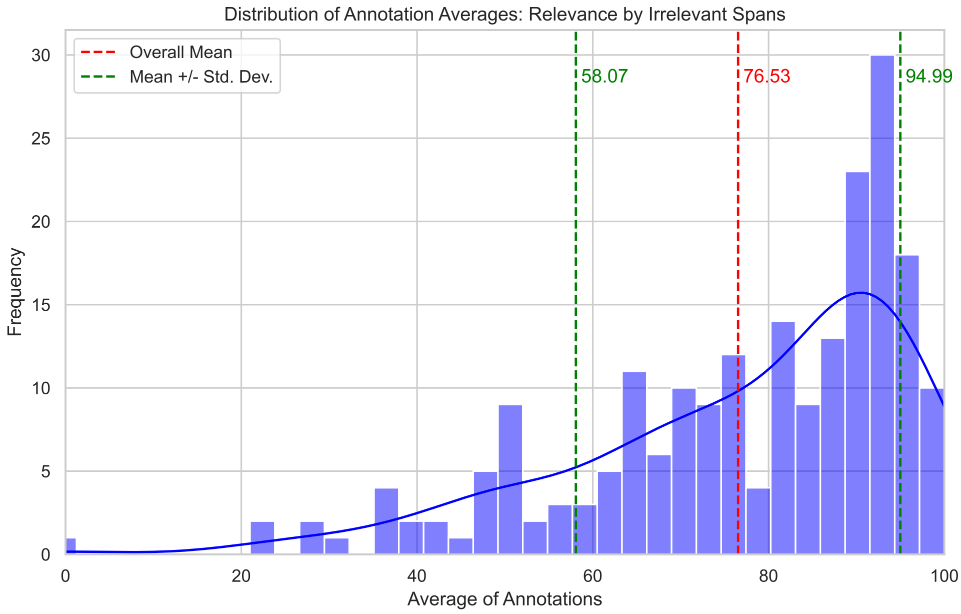


Figure 5.7 – Mean Distributions of Relevance score calculated by the irrelevant spans annotated by evaluators.
Source: The Author.

information storage, and natural language generation.

Analyzing score distributions and using the ICC, give different insights into how evaluators rate answers for completeness and relevance. Score distribution analysis shows how scores vary and their general pattern, while the ICC focuses on how much evaluators agree with each other. Interestingly, even though the scores for completeness are more uneven and spread out, evaluators agree more on these scores than on relevance scores. This means that while the scores are more spread out, they generally agree more on how complete an answer is.

In conclusion, while the relevance scores show a moderate degree of uniformity, completeness evaluations exhibit greater variability and some biases, as evidenced by the anomalies in the distribution. Such insights are important for understanding the evaluators' scoring behavior, which can inform the development of more effective QA systems that align with human judgment criteria.

5.4.3 Distributions of Irrelevant Spans Annotations

For a better understanding of relevance score annotations, Figure 5.7 shows the distribution of a relevance score calculated from the spans of texts marked as irrelevant

Evaluator	ICC3k
A	0.7939
B	0.9964
C	0.8672
D	0.9771

Table 5.3 – ICC3k correlation between evaluators' relevance score and their rate of the annotated relevant span of text.
Source: The Author.

by the annotators. The method for calculating the proportion of relevant parts within each answer text was established by the formula below, which quantifies the relevant portion of the text by subtracting the ratio of the annotated irrelevant span from the total length of the answer, providing a score for the relevance of the content.

$$\text{Proportion of Relevance} = 1 - \frac{\text{Length of Irrelevant Span}}{\text{Total Length of the Answer}}$$

The average of the relevant sections was 76.53, with a standard deviation of 18.46. This distribution is similar to the distribution of relevance scores shown in Figure 5.5, indicating a direct link between the amount of text marked as irrelevant and the relevance scores given by annotators. The matching average and standard deviation of these scores reveal that annotators consistently relied on marking irrelevant text to determine the relevance of an answer. Evaluators tend to give relevance scores that reflect the proportions of relevant text, emphasizing the significance of marking in the assessment process. This highlights the careful approach of the annotation process.

Table 5.3 presents each evaluator's ICC3k (A, B, C, and D) about the proportions of relevant parts of the answer annotation and the direct relevance score assigned by the evaluator. In other words, the ICC3k value shows how much the annotation of irrelevant parts made by the annotators agrees with the relevance score directly assigned by them to the answer. The ICC3k was used once there was a fixed set of raters evaluating all instances.

Evaluator B has the highest ICC3k value at 0.9964, which suggests almost perfect consistency in their relevance scoring compared to the span of text they marked as relevant. Evaluator D also shows a very high level of consistency with an ICC3k value of 0.9771.

Evaluators A and C, with ICC3k values of 0.7939 and 0.8672 respectively, exhibit a good level of agreement, though not as high as Evaluators B and D. These values are still relatively strong, indicating that while there may be some variability in their assessment of relevance, they are largely consistent.

Overall, the ICC3k results suggest a high level of agreement regarding the span of text identified as relevant and the direct relevance score. This suggests that the methodology for evaluating relevance based on the proportion of text not marked as irrelevant is a reliable measure for these evaluators. The high correlation reveals again that annotators consistently relied on marking irrelevant text to determine the relevance of an answer.

5.4.4 Comparison of GPT-4 and Reddit User Answers

The analysis comparing the annotation scores for answers generated by GPT-4 and Reddit users aims to evaluate the performance of these two sources in providing relevant and complete answers. This comparison is distinct from the prior analyses because it directly contrasts human answers with AI answers, focusing on determining which source produces better results according to human evaluators.

In the Table 5.4 and Table 5.5, “Wins” refers to the number of times the answers from a particular model (GPT-4 or Reddit User) were rated higher by an evaluator, while “Draw” indicates the number of times both models received the same score. Other values in the tables, such as “Mean”, “Median”, and “std”, provide statistical insights into the scores given by each evaluator for each criterion, highlighting the central tendency and dispersion of the scores.

Model	Evaluator	Mean	Median	std	Wins	Draw
GPT-4	A	95.5	100.0	13.3	64	0
	B	80.1	85.0	17.0	84	3
	C	78.2	85.0	21.3	77	24
	D	95.3	100.0	10.3	59	38
	Average	87.3	89.1	8.4	91	0
Reddit User	A	66.6	80.5	36.8	42	0
	B	72.4	80.5	26.2	19	3
	C	63.9	66.0	22.5	5	24
	D	52.2	56.5	39.8	9	38
	Average	63.8	64.1	19.7	15	0

Table 5.4 – Relevance
Source: The Author.

The results for the Relevance criterion in Table 5.4 show that GPT-4’s answers were generally considered more relevant by the evaluators, with higher mean and median scores compared to those of Reddit users. The standard deviation values for GPT-4 are

Model	Evaluator	Mean	Median	std	Wins	Draw
GPT-4	A	93.7	100.0	13.8	93	0
	B	67.3	72.0	25.9	83	4
	C	80.2	86.0	21.9	87	8
	D	71.7	87.0	31.0	87	13
	Average	78.2	85.1	19.0	103	0
Reddit User	A	59.4	65.0	31.9	13	0
	B	48.6	51.0	24.4	19	4
	C	68.1	72.0	22.0	11	8
	D	43.2	39.0	28.0	6	13
	Average	54.8	56.5	18.3	3	0

Table 5.5 – Completeness
Source: The Author.

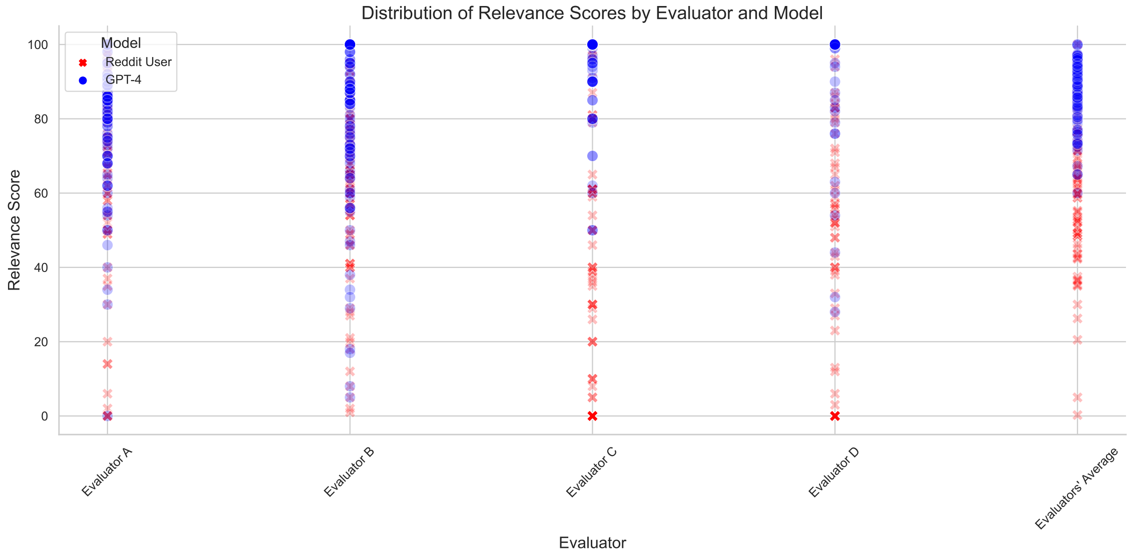


Figure 5.8 – Distribution of Relevance Scores by Evaluator for GPT-4 and Reddit User.
Source: The Author.

lower, indicating that the relevance scores for its answers were more consistent among different evaluators. For the Completeness criterion in Table 5.5, GPT-4 again outperforms Reddit users with higher mean and median scores and more wins, suggesting that GPT-4's answers were not only more relevant but also considered more complete by the evaluators.

The plots (Figures 5.8 and 5.9) visually represent the distribution of scores for each criterion by evaluator and model. These plots allow for a quick comparison of how often each source won in terms of higher scores and how the scores are distributed across different evaluators.

The superiority of GPT-4's answers in terms of relevance and completeness could be attributed to several factors. GPT-4 has been trained on a vast corpus of text, enabling

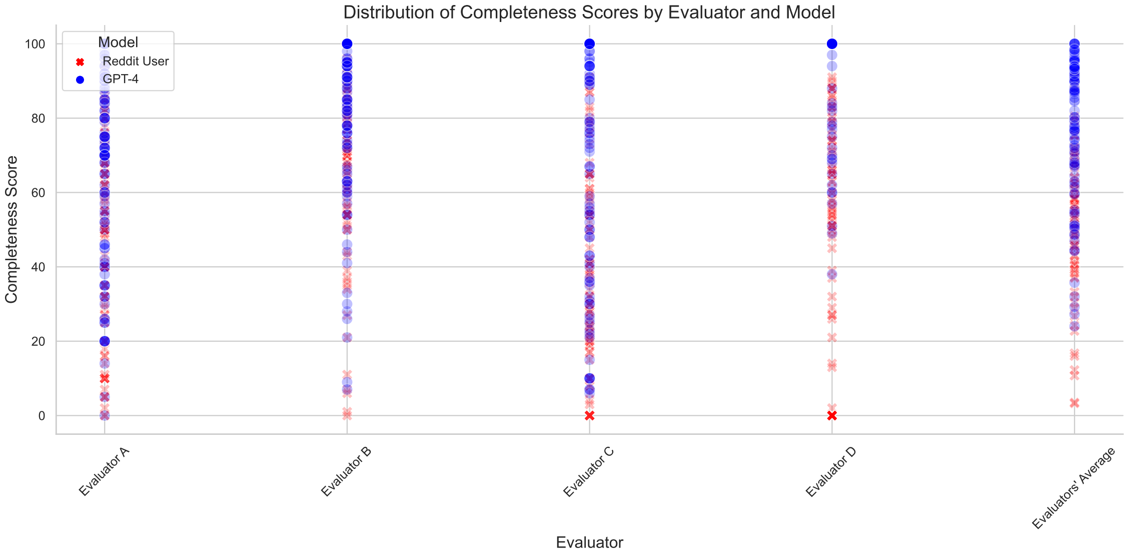


Figure 5.9 – Mean Distributions of Completeness Scores by Evaluator for GPT-4 and Reddit User.
Source: The Author.

it to generate answers that are not only contextually appropriate but also rich in content. In addition, the prompt used for GPT-4 was designed to control the size of the answers, ensuring they were comparable in length to the Reddit users' answers. Therefore, despite the constraint of matching the word count of Reddit users' answers, GPT-4 can effectively condense relevant information into concise answers. The AI's skill in considering many different sources can help make its answers more complete and relevant. Overall, the analysis suggests that GPT-4 is capable of producing answers that are more aligned with the human evaluators' expectations of relevance and completeness.

6 PROPOSED METRICS

This chapter is dedicated to describing the metrics proposed to evaluate completeness and relevance scores in long answers provided by non-factoid QA systems. The selection of these strategies was inspired by the literature review presented in Chapters 2 and 3, focusing mainly on the overlap of information units, such as n-grams, the training of regression models, and the use of prompt strategies with generative LLMs.

The metric models were evaluated on a dataset specifically constructed for this study. It is important to highlight that the metrics were developed independently of any analysis of the annotations present in the test set, ensuring the absence of data leakage during the design of the proposed metrics. In other words, data from this dataset was not used during the construction of the proposed approaches. Additionally, for metrics that rely on the training of supervised models, no instance or data derived from the test set was used. Thus, the results achieved can be considered reliable for generalization to new test sets.

6.1 Prompt Approach with Generative LLMs

The strategy involving the use of prompts and generative LLMs consists of formulating a prompt that explains to the LLM the concept of completeness or relevance metrics, where the model is required to evaluates an answer given to a specific question. The generative LLM employed in this approach must be previously trained to understand and execute instructions, as is the case with ChatGPT[1], which has the ability to answer a variety of instructions. Therefore, the instruction given to the model would be to assign a scoring value for the completeness or relevance metric.

Recent studies in the literature have demonstrated that the approach of prompts and LLMs generates significant results in tasks of assigning scores for evaluating text generated (GAO et al., 2024). For example, studies like (TöRNBERG, 2023), (OSTYAKOVA et al., 2023), and (CEGIN; SIMKO; BRUSILOVSKY, 2023) show that generative LLM achieves similar performance to crowdsourced annotators in various tasks related to score annotation. Similarly, the research of (KOCMI; FEDERMANN, 2023b) reveals that GPT-3.5 and GPT-4 achieved the best results in benchmarks for the accuracy of translation quality assessment, overcoming the performance of all other models from the WMT22

[1] https://chat.openai.com/

metric shard task and with the considerable correlation with to human judgments. For these reasons, one of the approaches proposed and evaluated in this study is implementing a prompt strategy for assigning relevance and completeness scores.

Intuitively, the task of assigning completeness and relevance scores to long answers requires a certain level of world knowledge and judgment capability, given that determining which components of the answer are relevant or identifying what information is missing necessitates specific knowledge about the question subject. Moreover, the definition of what is relevant can vary among individuals, introducing a characteristic of subjectivity that demands an ability to form opinions. Considering these factors, the approach that uses prompts for generative LLMs shows promise, as some of these models have a remarkable ability to understand real world context and formulate opinions about differents subjects.

In the context of this study, the use of the GPT-3.5 and GPT-4 LLMs was chosen through API calls, setting the temperature parameter to zero. This means that the model will generate more deterministic and consistent predictions, always opting for the next most likely word in its training, without introducing random variations. These two models were tested individually, applying the same test set and the same prompt strategy. The choice of these models is due to their superior performance in similar tasks, compared to other LLMs specialized in following instructions (LIU et al., 2023; KOCMI; FEDERMANN, 2023a).

The prompt strategy used in this study requires the use of a significant number of tokens, especially for lengthy answers, which can exceed 3,000 tokens (according to the tokenization model used by OpenAI[2]) for a single instance. This justifies the choice of the GPT-3.5 and GPT-4 models, as smaller models may not support large sized prompts due to max tokens limitations. Additionally, the financial aspect must be considered, as the use of these models implies significant costs, making it unfeasible to use a wider variety of LLMs for the purpose of this research.

The prompts developed for this task are employed individually for the completeness and relevance metrics. In summary, the prompt based approach involves explaining the concept of completeness, relevance, and accuracy metrics in detail, clarifying that the goal is to assign a completeness or relevance score. The model is requested to examine an answer and identify which information is missing, in the context of completeness, or to determine which segments of the text are relevant or irrelevant, in the case of relevance.

[2] https://platform.openai.com/tokenizer

Finally, a numerical score from 0 to 100 is requested.

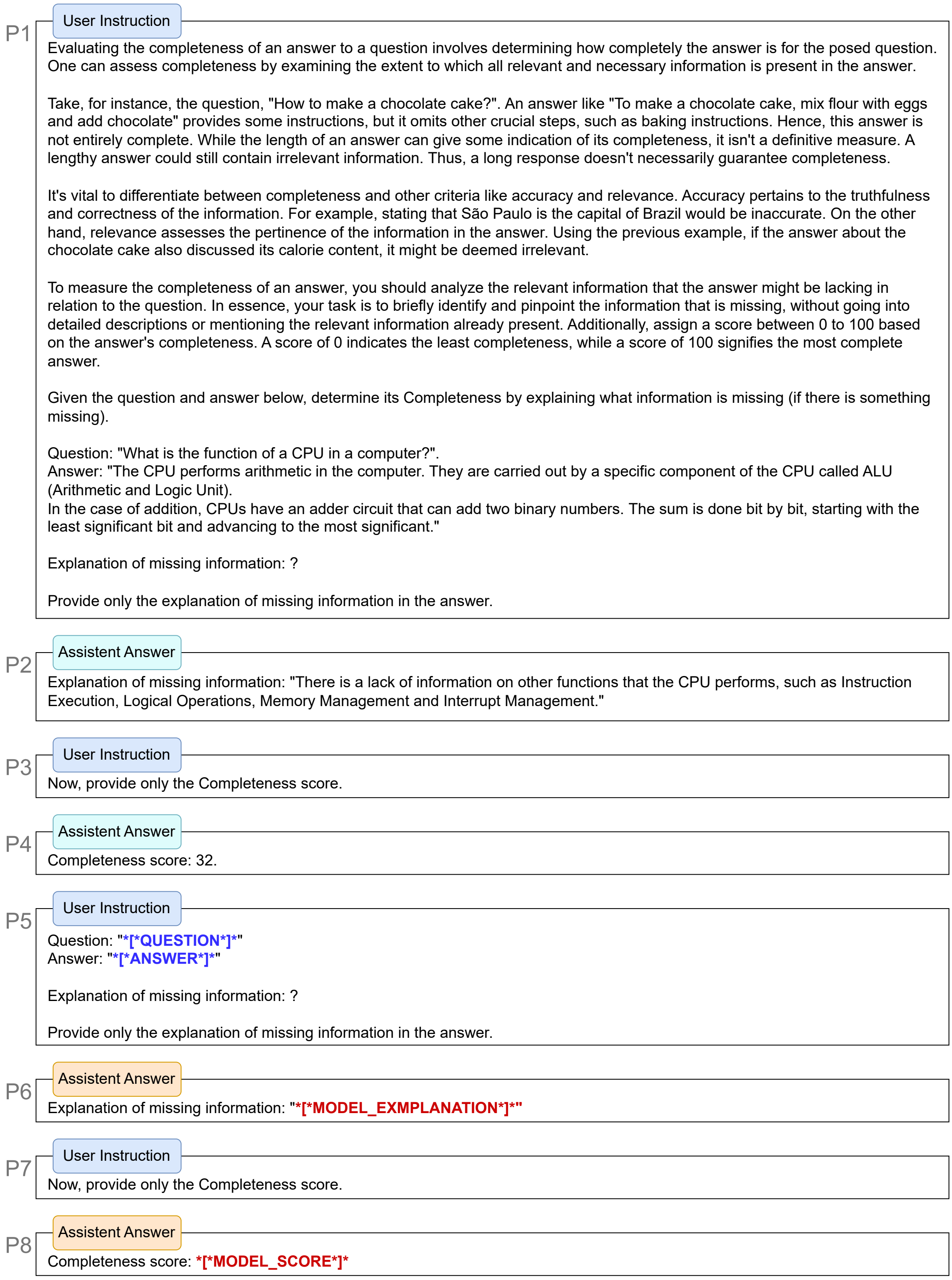
6.1.1 Completeness Prompt

Figure 6.1 demonstrates the prompt methodology used for determining the completeness score. This strategy is divided into eight phases, as illustrated in the figure, where the sections marked as “User Instruction” represent the guidelines provided by the user, and “Assistant Answer”, the answers formulated by the model. It is important to highlight that phases P2 and P4 are not outputs generated by the model but manually created examples to exemplify the expected model output. These examples are integrated into the model through the structure supported by the OpenAI API for GPTs. Only the answers from phases P6 and P8 are directly produced by the model.

The initial part (P1) is dedicated to explaining the process of scoring for completeness. Initially, the task is contextualized through a clear and objective description accompanied by a practical example, aiming to establish the understanding of the objective. Subsequently, a distinction is made between the criterion of completeness and the criteria of relevance and accuracy, with the intent to prevent possible confusion by the model regarding these criterias. The explanation then proceeds on how the completeness of an answer can be measured, suggesting that the model evaluates which essential components, related to the question, are missing in the answer. In other words, the strategy aims to instruct the model to first identify what relevant elements are missing in the answer.

To prevent the model from generating additional information that is unnecessary for the task, such as identifying parts it considers relevant, an instruction is explicitly included in the prompt to concentrate on identifying what is missing. Moreover, P1 clarifies the assignment of the score for the answer, which ranges from 0 (completely incomplete) to 100 (fully complete).

The conclusion of the first part presents a detailed example, illustrating how the model should proceed in assigning the completeness score. In this exemple, through a provided question and its respective answer, the initial task of the model consists of detailing which information is missing in this answer. To prevent the generation of unnecessary content, an explicit request is made for the model to limit itself to providing only the explanation of the missing information. This example is used for all entries and is selected for belonging to the domain of the evaluated answers, in this case, computer



P1
User Instruction

Evaluating the completeness of an answer to a question involves determining how completely the answer is for the posed question. One can assess completeness by examining the extent to which all relevant and necessary information is present in the answer.

Take, for instance, the question, "How to make a chocolate cake?". An answer like "To make a chocolate cake, mix flour with eggs and add chocolate" provides some instructions, but it omits other crucial steps, such as baking instructions. Hence, this answer is not entirely complete. While the length of an answer can give some indication of its completeness, it isn't a definitive measure. A lengthy answer could still contain irrelevant information. Thus, a long response doesn't necessarily guarantee completeness.

It's vital to differentiate between completeness and other criteria like accuracy and relevance. Accuracy pertains to the truthfulness and correctness of the information. For example, stating that São Paulo is the capital of Brazil would be inaccurate. On the other hand, relevance assesses the pertinence of the information in the answer. Using the previous example, if the answer about the chocolate cake also discussed its calorie content, it might be deemed irrelevant.

To measure the completeness of an answer, you should analyze the relevant information that the answer might be lacking in relation to the question. In essence, your task is to briefly identify and pinpoint the information that is missing, without going into detailed descriptions or mentioning the relevant information already present. Additionally, assign a score between 0 to 100 based on the answer's completeness. A score of 0 indicates the least completeness, while a score of 100 signifies the most complete answer.

Given the question and answer below, determine its Completeness by explaining what information is missing (if there is something missing).

Question: "What is the function of a CPU in a computer?".
Answer: "The CPU performs arithmetic in the computer. They are carried out by a specific component of the CPU called ALU (Arithmetic and Logic Unit).
In the case of addition, CPUs have an adder circuit that can add two binary numbers. The sum is done bit by bit, starting with the least significant bit and advancing to the most significant."

Explanation of missing information: ?

Provide only the explanation of missing information in the answer.

Explanation of missing information: "There is a lack of information on other functions that the CPU performs, such as Instruction Execution, Logical Operations, Memory Management and Interrupt Management."

Now, provide only the Completeness score.

Completeness score: 32.

Question: "*[*QUESTION*]*"
Answer: "*[*ANSWER*]*"

Explanation of missing information: ?

Provide only the explanation of missing information in the answer.

Explanation of missing information: "*[*MODEL_EXMPLANATION*]*"

Now, provide only the Completeness score.

Completeness score: *[*MODEL_SCORE*]*

Figure 6.1 – Prompt for assigning the completeness score by GPT models.
Source: The Author.

science. Also, the example was chosen for clearly demonstrating a case of an answer lacking essential information.

The second part (P2) displays an example of output that the model should generate when requested with real data, which is explanations about the missing data in a given answer. In this context, it is expected that this information will help the model differentiate which information is missing when assigning the final completeness score. For example, a detailed and extensive explanation, indicating the absence of multiple relevant elements, suggests the need to assign a low completeness score to the answer. In other hand, an explanation that does not identify absence of relevant elements indicates that the answer has a high degree of completeness.

In the third part (P3), the model is required to determine and assign a completeness score to the analyzed answer. Subsequently, in part four (P4), a practical example of the completeness score for a specific answer is presented to the model. It is important to emphasize that, although the selected example aims to be clear and minimally subjective regarding the assigned score, the model is expected to make its evaluations based on the criteria illustrated in the example. It is important to note that an example with a low score was chosen, given that previous empirical observation with ChatGPT indicated a tendency of the model to provide higher evaluations to the answers. Thus, presenting an example with a low score seeks to show the model the existence of answers with low scores. The inclusion of additional examples is a possibility, although this would result in an increase in the size of the prompt and, consequently, in the cost of using this strategy, especially considering the use of API calls.

In section (P5), the analysis of the real data input by the model begins, which consists of a question and an answer, represented by "*[*QUESTION*]*" and "*[*ANSWER*]*", respectively. Therefore, an explanation of the missing data for the actual data to be evaluated is requested. The following stage, (P6), consists of an output produced by the model, in which it must provide a detailed explanation about the information missing in the provided answer. This explanation is intended to assist in assigning the final completeness score by the model.

Part (P7) is used to requesting the final completeness score from the model. The value corresponding to this score is revealed in section (P8), where "*[*MODEL_SCORE*]*" represents the completeness score determined by the model after its analysis for the real data.

6.1.2 Relevance Prompt

The prompt strategy for evaluating the relevance criterion shares similarities with the approach used for analyzing completeness, being structured into eight distinct parts, as illustrated in Figure 6.2. In the relevance assessment, part P1 is dedicated to providing a detailed description and specific examples that demonstrate the concept of relevance. In this way, the example presented in P1 shows an extensive answer, full of irrelevant information, which, although large, serves as an example of an answer with low relevance.

Another distinction related to completeness is that the prompt asks the model to analyze the answer to identify the relevant and irrelevant parts. Thus, the model can adopt a logic based on the proportion of relevant information in contrast to what is not. For example, if 75% of the answer is considered relevant and 25% irrelevant, the relevance score could be set at 75. In this way, in section P2, an example of relevant and irrelevant information is presented, instructing the model to extract sentences directly from the answer and put them to the fields corresponding to the relevant and irrelevant information. Consequently, sections P3 and P4 request a relevance score from the model, illustrating it with a specific scoring example.

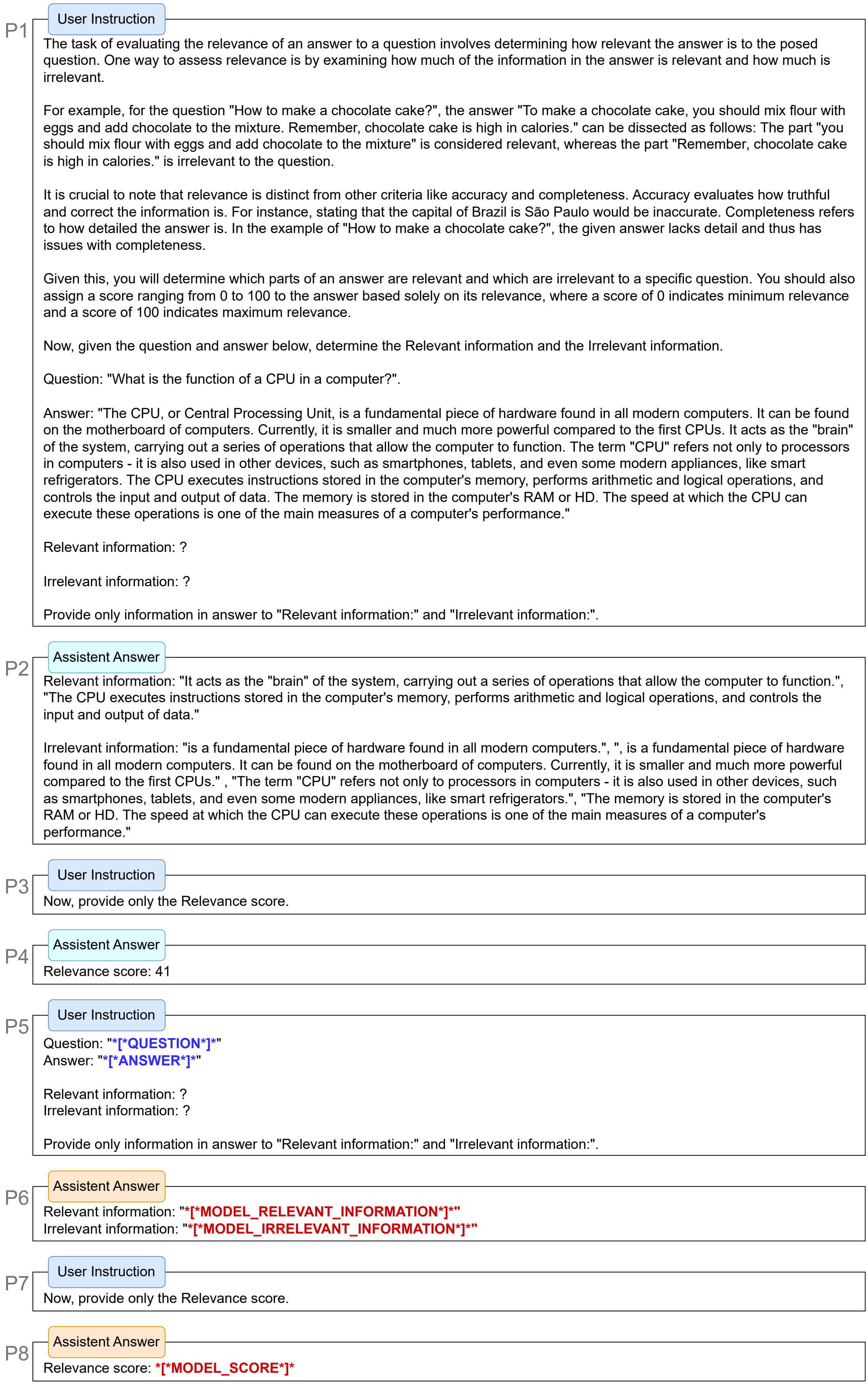
Following this approach, sections P5, P6, P7, and P8 adopt the same logic for the real data. However, in these stages, the model is responsible for generating the relevant and irrelevant information, presented in section P6. Based on this information, a relevance score is requested in section P7, which is revealed in section P8, replacing the placeholder “*[*MODEL_SCORE*]*” with the corresponding value for the real data.

6.2 Information Units

The completeness and relevance criteria are directly linked to recall and precision functions. Recall is a metric that quantifies the proportion of all relevant units correctly identified in relation to the total existing relevant units. On the other hand, precision represents the proportion of identified units that are truly relevant in relation to the total units identified:

$$\text{Recall} = \frac{\text{Number of relevant units identified}}{\text{Total existing relevant units}}$$

$$\text{Precision} = \frac{\text{Number of relevant units identified}}{\text{Total units identified}}$$



P1 User Instruction

The task of evaluating the relevance of an answer to a question involves determining how relevant the answer is to the posed question. One way to assess relevance is by examining how much of the information in the answer is relevant and how much is irrelevant.

For example, for the question "How to make a chocolate cake?", the answer "To make a chocolate cake, you should mix flour with eggs and add chocolate to the mixture. Remember, chocolate cake is high in calories." can be dissected as follows: The part "you should mix flour with eggs and add chocolate to the mixture" is considered relevant, whereas the part "Remember, chocolate cake is high in calories." is irrelevant to the question.

It is crucial to note that relevance is distinct from other criteria like accuracy and completeness. Accuracy evaluates how truthful and correct the information is. For instance, stating that the capital of Brazil is São Paulo would be inaccurate. Completeness refers to how detailed the answer is. In the example of "How to make a chocolate cake?", the given answer lacks detail and thus has issues with completeness.

Given this, you will determine which parts of an answer are relevant and which are irrelevant to a specific question. You should also assign a score ranging from 0 to 100 to the answer based solely on its relevance, where a score of 0 indicates minimum relevance and a score of 100 indicates maximum relevance.

Now, given the question and answer below, determine the Relevant information and the Irrelevant information.

Question: "What is the function of a CPU in a computer?".

Answer: "The CPU, or Central Processing Unit, is a fundamental piece of hardware found in all modern computers. It can be found on the motherboard of computers. Currently, it is smaller and much more powerful compared to the first CPUs. It acts as the "brain" of the system, carrying out a series of operations that allow the computer to function. The term "CPU" refers not only to processors in computers - it is also used in other devices, such as smartphones, tablets, and even some modern appliances, like smart refrigerators. The CPU executes instructions stored in the computer's memory, performs arithmetic and logical operations, and controls the input and output of data. The memory is stored in the computer's RAM or HD. The speed at which the CPU can execute these operations is one of the main measures of a computer's performance."

Relevant information: ?

Irrelevant information: ?

Provide only information in answer to "Relevant information:" and "Irrelevant information:".

P2 Assistent Answer

Relevant information: "It acts as the "brain" of the system, carrying out a series of operations that allow the computer to function.", "The CPU executes instructions stored in the computer's memory, performs arithmetic and logical operations, and controls the input and output of data."

Irrelevant information: "is a fundamental piece of hardware found in all modern computers.", ", is a fundamental piece of hardware found in all modern computers. It can be found on the motherboard of computers. Currently, it is smaller and much more powerful compared to the first CPUs." , "The term "CPU" refers not only to processors in computers - it is also used in other devices, such as smartphones, tablets, and even some modern appliances, like smart refrigerators.", "The memory is stored in the computer's RAM or HD. The speed at which the CPU can execute these operations is one of the main measures of a computer's performance."

P3 User Instruction

Now, provide only the Relevance score.

P4 Assistent Answer

Relevance score: 41

P5 User Instruction

Question: "***[*QUESTION*]***"
Answer: "***[*ANSWER*]***"

Relevant information: ?
Irrelevant information: ?

Provide only information in answer to "Relevant information:" and "Irrelevant information:".

P6 Assistent Answer

Relevant information: "***[*MODEL_RELEVANT_INFORMATION*]*"**
Irrelevant information: "***[*MODEL_IRRELEVANT_INFORMATION*]*"**

P7 User Instruction

Now, provide only the Relevance score.

P8 Assistent Answer

Relevance score: ***[*MODEL_SCORE*]***

Figure 6.2 – Prompt for assigning the relevance score by GPT models.
Source: The Author.

The proposed approach is based on these concepts through metrics based on the classic notions of recall and precision in information retrieval systems. The completeness of an answer is determined similarly to recall. For a specific question, there exists a defined set of all relevant information that the answer should cover. This set corresponds to the *Total existing relevant units*. For a given answer, there is a subset containing the relevant information present in that answer. This subset corresponds to the *Number of relevant units identified*. Therefore, the completeness of an answer can be understood as the proportion of relevant information present in the answer in relation to the total set of relevant information that the answer should contain.

Relevance have a direct connection with precision. In this context, for a given question, the total set of information present in the answer, which is related to the *Total units identified*, is considered. Within this total set, there is a subset of information that is relevant to the answer and, consequently, a subset of information that is not relevant. Thus, the set of relevant information corresponds to the *Number of relevant units identified*. Therefore, the relevance of an answer can be interpreted as the proportion of relevant information found in the answer divided by the total amount of information present in the answer.

The evaluation of the completeness and relevance of answers involves quantifying how much an answer have the necessary information (completeness) and how precise and relevant that information is (relevance). To do this, sets of information units are defined that model the answers and the relevant information, illustrated in Figure 6.3. This figure represents a universe of information units for a specific answer, identifying a subset of relevant and irrelevant units. The circle symbolizes the area containing the information present in the answer, which includes both relevant and irrelevant units. The set of information units of an answer is defined exclusively by the information contained in the text of the answer. Similarly, the set of relevant information is equally limited, covering only the information that is indeed relevant to the answer. On the other hand, the set of irrelevant units comprises any information that does not contribute to the relevance of the answer. It is important to consider that the definition of what constitutes relevant information can vary according to the perspective of each person.

First, the set of all relevant information units (R_{all}) is defined as a universal set of information relevant to the input question. The answer to be evaluated is then represented by a set of information units (U_{answer}), which includes both relevant (R_{answer}) and irrelevant information. The challenge is to quantify how complete and relevant the provided

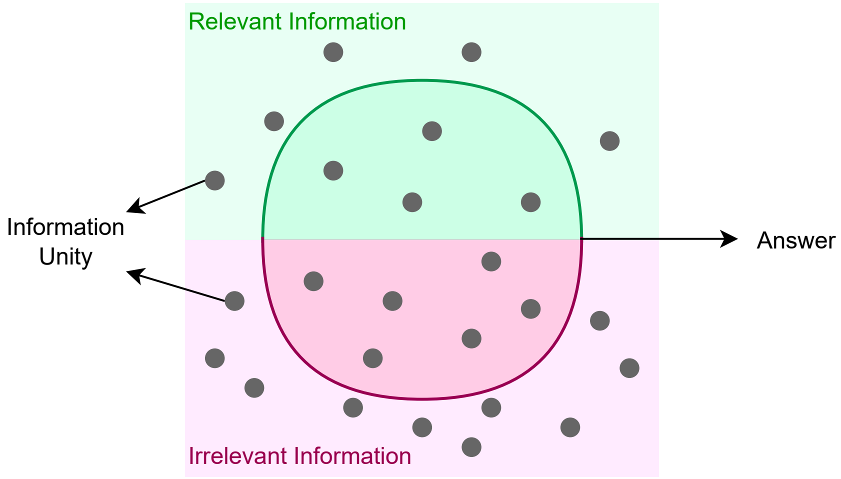


Figure 6.3 – Information unities of an answer.
Source: The Author.

answer is.

Completeness is evaluated by the ratio between the relevant information present in the answer and the total set of relevant information:

$$\text{Completeness} = \frac{|R_{answer}|}{|R_{all}|}$$

where $|R_{answer}|$ represents the cardinality of the set of relevant information units present in the answer, and $|R_{all}|$ is the cardinality of the total set of relevant information.

Relevance is determined by the proportion of information in the answer that is relevant:

$$\text{Relevance} = \frac{|R_{answer}|}{|U_{answer}|}$$

where $|U_{answer}|$ indicates the cardinality of the set of all information units in the answer.

Thus, R_{answer} and U_{answer} are defined as follows:

- $R_{answer} = U_{answer} \cap R_{all}$, symbolizing the intersection between the information units in the answer and the set of all relevant information units.
- $U_{answer} = S(t_{answer})$, where $S(\cdot)$ is a function that segments the text of the answer (t_{answer}) into discrete information units.
- Similarly, $R_{all} = S(t_{ref})$, where t_{ref} is a reference text containing all the information considered relevant to the context under analysis.

Considering this formal definition, challenges related to the implementation of this method arise, including: the representation of information units; the segmentation of a text into information units ($S(\cdot)$); the definition of the reference text t_{ref} for the answers;

determining the relevance of a unit; and, finally, the application of formulas analogous to recall and precision.

6.2.1 Segment information units

To segment the text of answers or reference contents, the technique of dividing into sentences is employed. The premise is that by dividing the text into sentences, each information unit corresponds to a complete expression of an idea, avoiding being as broad as a paragraph, but also not as limited as words or short phrases. Compared to more complex approaches that require deep semantic analysis, such as efficient segmentation into triples, sentence division presents itself as a relatively simple task in NLP, ensuring efficiency and accessibility to the process. Additionally, the flexibility of this technique, which can be applied to any type of text, regardless of the domain or complexity of the subject, is highlighted.

Besides sentence segmentation, there are alternative methods for dividing a text into information units, such as the use of more restricted units, such as words and n-grams. However, these units, when used in isolation, may not capture a complete idea, leading to a loss of context and semantic cohesion. Other approaches, like the use of tuples and triples (subject, predicate, object), offer greater semantic richness, however, representing these structures through vectors of embeddings could faces significant challenges. Although sentence embeddings are designed to encompass the global semantics of a text unit, including contextual and relational details, the representation of tuples and triples may not be as effective in this format. This is due to the fact that sentence embeddings are optimized to conserve semantic information in a continuous and dense dimension, reflecting the complexity and depth of meaning expressed in natural text. On the other hand, tuples and triples, due to their structured nature and focus on specific relations, may lose aspects of their relational and semantic context. In any case, the idea of using tuples and triples to represent units of information is interesting and should be a research problem in future research.

6.2.2 Represent units of information

After segmenting the text of the answer and the reference contents into sentences, the information units are represented through embedding vectors that are sensitive to the context, using models such as BERT. This strategy is chosen, once the methodology aligns with the need to capture the semantic complexity and contextual details of the information units, facilitating a richer and multidimensional representation of the textual content.

Language models based on BERT are especially suitable for this task (DEVLIN et al., 2018), as they were developed with the aim of understanding the context of each word within a sentence. This understanding tke into account the interaction among all words present. As a result, the generated embeddings reflect the complete semantics of the information unit, being capable of encapsulating the meaning of a sentence.

The representation of a sentence in information units is carried out in the process that begins with a sentence $S = \{w_1, w_2, ..., w_n\}$ composed of n words. The BERT model maps each word w_i to an embedding vector v_i in a d-dimensional feature space, taking into account the context provided by adjacent words. Therefore, for the word w_i, its embedding vector is obtained through:

$$v_i = BERT(w_i; \{w_1, ..., w_{i-1}, w_{i+1}, ..., w_n\})$$

where $v_i \in R^d$ is the embedding vector for the word w_i, and the function $BERT(\cdot)$ represents the embedding mechanism of the model, which considers both the target word and its context. To represent the complete sentence S as a single embedding vector, the vectors are aggregated through their mean:

$$V_S = \frac{1}{n}\sum_{i=1}^{n} v_i$$

where $V_S \in R^d$ is the resulting embedding vector for the sentence S, represented by the average of the semantic features of all the words contained in the sentence. The BERT model supports up to 512 tokens in an input. In cases where the sentence contains more tokens than this, only the first tokens are used, and the last are disregarded.

The semantic representation of a complete answer or reference content through information units is done with an embedding matrix. Each sentence of the answer or information unity is transformed into an embedding vector V_S, which composes a row of this matrix. Thus, if an answer or reference content is composed of m sentences, the total semantic representation will be a matrix M with dimensions $m \times d$, where d is the dimension of the feature space of the embeddings.

6.2.3 Reference set

Ideally, the reference set should be an answer containing exclusively the information relevant to the individual who posed the question. Developing this type of reference resource can be challenging, especially for extensive answers that require the inclusion of detailed explanations, in-depth analyses, instructions, and other elements that specifically serve the user who asked the question. In a test dataset that includes a reference answer, this can be used to define the reference set. However, in the absence of a reference answer in the test set, another resource is needed to serve as a universal set of information pertinent to the input question (R_{all}).

Search engines are tools that facilitate the search for documents containing information relevant to a set of keywords. Therefore, in situations where no reference answers are available, the contents obtained through search engines, using keywords extracted from the input question, can be employed as reference content. However, this type of approach can present challenges related to the processing of such information, given that the content found may include various irrelevant information. Web pages, for example, can contain elements not related to the searched subject, such as advertisements and links to other contents that are irrelevant.

It is important to highlight that the technical development of this model took into account the dataset used for testing only to avoid the use of data that derived from the test dataset, such as Reddit pages. For the proposed strategy with information units, three types of reference content are suggested:

- The **reference answer** to the question available in the test set.
- **Web pages** returned by the Google search engine, using the input question as the query. For the extraction of pages, the Google API was used, searching, for each question, the top ten English pages ranked highest. Reddit pages were excluded, since the input question from test dataset was extracted from this forum. No complex processing was performed on the pages, only the removal of HTML tags.
- The **Wikipedia** text that most closely matches the content of the input question. For this, the topics of the question were first extracted, focusing mainly on nouns

and excluding stop words. Then, the Python “wikipedia” library was used to search for the English page most relevant to the question’s topic.

By proposing three distinct approaches to reference content, it becomes feasible to evaluate which one presents the best performance in different situations. This way, it is possible to investigate, for example, if the approaches that do not use a reference answer offer satisfactory performance for determining the completeness and relevance of the answers.

6.2.4 Determine relevance of a unit of information

Once the text of the answer to be evaluated is segmented into pieces and then transformed into information units, it is necessary to determine which units are relevant and which are irrelevant. In this proposal, the text is segmented into sentences and then transformed into embedding vectors. To determine whether a particular embedding vector of the answer is relevant or not, cosine similarity between the embedding vectors of the answer and the embedding vectors of the reference content is used. Let $V_{S_{answer}}$ be the set of embedding vectors of the information units of the answer and $V_{S_{ref}}$ the set of embedding vectors of the information units of the reference content. The relevance of each information unit in the answer is determined by the maximum cosine similarity between that unit and the units of the reference content. Thus, the cosine similarity between two vectors a and b is given by:

$$\text{cosine\_similarity}(a, b) = \frac{a \cdot b}{||a||||b||}$$

where $a \cdot b$ is the dot product of vectors a and b, and $||a||$, $||b||$ are the Euclidean norms of vectors a and b, respectively.

For each embedding vector $v_{answer} \in V_{S_{answer}}$, the maximum cosine similarity with the embedding vectors of the reference content $V_{S_{ref}}$ is calculated:

$$\text{max\_similarity}(v_{answer}, V_{S_{ref}}) = \max_{v_{ref} \in V_{S_{ref}}} \text{cosine\_similarity}(v_{answer}, v_{ref})$$

If max_similarity$(v_{answer}, V_{S_{ref}})$ is greater than a predefined threshold θ, then the information unit represented by v_{answer} is considered relevant; otherwise, it is considered irrelevant. This process is repeated for all information units of the answer to be evaluated, resulting in the determination of the set R_{answer}, which contains all information units considered relevant within the answer.

Thus, the determination of the relevance of an information unit is defined as:

$$R_{answer} = \{v_{answer} \in V_{S_{answer}} | \text{max\_similarity}(v_{answer}, V_{S_{ref}}) > \theta\}$$

This approach allows for the quantitative determination of the amount of relevant units of the answer in relation to the reference content, providing the basis for the analysis of completeness and relevance. Also, this approach defines the intersection between the set of units of the answer and the set of all relevant units, represented by $U_{answer} \cap R_{all}$.

6.2.5 Application of formulas

After segmenting the textual contents and transforming the segments into information units, in addition to determining the units relevant to the answer, it is feasible to calculate the completeness and relevance of the information. To do this, the cardinality of the set of relevant units present in the answer ($|R_{answer}|$), and the cardinality of the set containing all relevant units ($|R_{all}|$) are used to assess completeness. To measure relevance, the cardinality of the set that encompasses all units of the answer ($|U_{answer}|$) is employed. Completeness and relevance are then determined in a manner analogous to their respective definitions.

6.3 Strategy with Regression Model

The strategy using a regression model consists of training a supervised model with an input of a question and an answer, and an output being the score of completeness or relevance. Then, from this trained model, it's possible to make predictions of these criteria using a question and an answer as input, without the need for reference answers.

In this approach, an individual BERT model is used for each criterion (completeness and relevance) as the basis for training, since it is a pre-trained language model that establishes itself as a standard in NLP tasks due to its training on extensive textual datasets. This process adjusts the model's parameters to capture complex language de-

```
Question: *[*QUESTION*]*\n\n Answer: *[*ANSWER*]*\n\nHow *[*CRITERIA*]* is this answer?
```

Figure 6.4 – Input format for regression model.
Source: The Author.

tails, enabling it to understand contexts and textual meanings. When refined for specific tasks, BERT transfers its comprehensive linguistic knowledge, improving performance in different applications, from text comprehension to language generation, providing a solid foundation for customization in particular text processing needs.

Other language models, similar to BERT, that can also be applied to the proposed approach include RoBERTa (LIU et al., 2019), ALBERT (LAN et al., 2020), DistilBERT (SANH et al., 2020), and ELECTRA (CLARK et al., 2020). However, BERT is chosen for being a pioneering and fundamental model for various researches, influencing the development of techniques and models in NLP subsequently. Therefore, it represents a well established and reliable starting point for the proposed approach with a regression model to predict the completeness and relevance of answers.

For the fine-tuning of the model, a synthetically created dataset was used, which is explained in the subsection below, containing 100,000 examples. The model's input is represented in the format shown in Figure 6.4, where the question is placed in place of *[*QUESTION*]*, the answer is placed in place of *[*ANSWER*]*, and for the criteria of completeness and relevance, *[*CRITERIA*]* is replaced by "complete" or "relevant" respectively. The model supports an input with a maximum of 512 tokens, considering all this information. The standard training values, justified by empirical observations, Based on standard training values, justified by empirical observations, the training settings were used: a batch size of 8 is used for training, along with the Adam optimizer, 5 epochs, and a learning rate of 0.00001.

6.3.1 Synthetic Dataset

The synthetic dataset is created using instances from the ELI5 dataset and (XU et al., 2023), which we referred to as WebGPT. Both datasets contains a large collection of questions with detailed answers. For ELI5, we carefully avoided using questions that appear in the test dataset used in experiments in this study. The ELI5 instances were extracted using a specific script, detailed in subsection 5.1 of this study. The WebGPT dataset includes over 15,000 instances, from which a small subset was selected for training the regression model. The purpose of utilizing two datasets is to diversify the style of questions and answers. Therefore, mitigating inherent biases in these datasets and enhancing the model's ability to generalize.

This dataset is called synthetic because the answers from ELI5 and WebGPT are modified to vary their degree of completeness and relevance. In this case, when the answer is kept original, its completeness and relevance score is maximum. Removing original content from an reference answer decreases its completeness proportionally. Similarly, including irrelevant information reduces its relevance score.

Synthetic answers are crafted by selecting, modifying, and integrating different text segments. Initially, segments from the original reference answer, which are considered totally relevant, are extracted. To these, segments from two additional sources are added: documents directly related to the question and random excerpts from Wikipedia. Each segment's relevance is assessed using cosine similarity with the question. Segments from the original answers are considered fully relevant. Segments from documents related to the question are given an intermediate importance based on cosine similarity, indicating their greater relevance compared to random sources, which receive the lowest relevance weight.

More specifically, the method for generating these synthetic answers can be divided into the following main steps:

1. **Initial Preparation:** The reference answer (ref_answer) and text from other sources are segmented into individual sentences. This division facilitates the manipulation and analysis of discrete units of information.
2. **Sentence Selection and Removal:** A specified percentage of sentences is removed from the reference answer to create an information gap, forming the basis for the synthetic answer. This process is governed by a removal rate ($remove\_rate$), which is dynamically adjusted to vary the extent of information omission.
3. **Inclusion of Random and Contextual Sentences:** Additional sentences are chosen from both a pool of random sentences ($rand\_sentences$) and from the text within the relevant document ($document\_sentences$). The inclusion of these sentences is dictated by predetermined inclusion quantities ($inclusion\_quantity$) for each source, aiming to enrich the synthetic answer with pertinent contextual information and diverse content.
4. **Combination:** The remaining sentences from the original reference answer and the selected additional sentences are merged to form the synthetic answer. This

combination strategy involves placing the original sentences at the beginning of the answer, while incorporating the additional sentences towards the end.

5. **Completeness and Relevance Evaluation:** Using vector representation techniques with the BERT model and cosine similarity measures, the completeness and relevance of the synthetic answer are assessed. This evaluation considers both the amount of information removed from the original answer and the volume of information added.

Through these steps, a synthetic answer is created to simulate a real answer to a question. This process involves adding or removing segments of information, which vary in relevance to the question. This method helps in generating a diverse dataset for training regression models that evaluate the completeness and relevance of answers.

The completeness scores of a synthetic answer are determined by the proportion of relevant information it contains relative to the reference answer. Therefore, completeness is a value between 0 and 100, calculated as follows:

$$\text{Completeness} = \frac{h_{completeness}(a_{ref}, u_{ref}, a_{missref}, u_{rand}, u_{alphas})}{|a_{ref}|}$$

Where:

- a_{ref} represents the information units in the reference answer. Thus, $|a_{ref}|$ is the cardinality of the set of information units in the reference answer.
- u_{ref} represents the information units of a_{ref} present in the synthetic answer.
- $a_{missref}$ are the units of a_{ref} that are missing in the synthetic answer.
- u_{rand} represents random information units included in the synthetic answer.
- u_{alphas} are the weights associated with each random unit u_{rand}, reflecting their importance or relevance.
- $h_{completeness}(\cdot)$ is a function that calculates a score based on the similarity between u_{rand} and $a_{missref}$, adjusted by the weights u_{alphas}, in addition to considering the units present both in the reference answer and in the synthetic answer.

The relevance score, on the other hand, measures how relevant the information provided by the synthetic answer is in relation to the question. It is defined by:

$$\text{Relevance} = \frac{h_{relevance}(a_{ref}, u_{ref}, a_{missref}, u_{rand}, u_{alphas})}{|a_{sys}|}$$

- a_{sys} represents the information units in the synthetic answer.
- $h_{relevance}(\cdot)$ is a function that calculates a score based on the similarity that emphasizes the relevance of the information included in relation to the question.
- The other terms have the same meaning as in the completeness formula.

The function $h_{completeness}(\cdot)$ considers the quantity and quality of the information present in the answer compared to the reference answer. Meanwhile, the function $h_{relevance}(\cdot)$ focuses on assessing how appropriately the information addresses the original question. These functions take into account the similarities between the randomly added information units and those missing in the reference answer, adjusted by the weights u_{alphas}, providing a measure of how well the synthetic answer comprehends the essential information of the reference answer and how relevant that information is to the subject of the question.

The function $h_{completeness}(\cdot)$ can be defined as:

$$h_{completeness}(a_{ref}, u_{ref}, a_{missref}, u_{rand}, u_{alphas}) = |u_{ref}| + norm\_s(u_{rand}, a_{missref}, u_{alphas})$$

Where:

- $|u_{ref}|$ is the quantity of information units from the reference answer present in the synthetic answer.
- $norm\_s(\cdot)$ is a function that normalizes the sum of similarities between the random information units (u_{rand}) and the missing units from the reference answer ($a_{missref}$), weighted by the weights (u_{alphas}). The normalization ensures that the contribution of the added information does not exceed the importance of the original information from the reference answer.

The function $h_{relevance}(\cdot)$ can be expressed as:

$$h_{relevance}(a_{ref}, u_{ref}, a_{missref}, u_{rand}, u_{alphas}) = |u_{ref}| + norm\_s(u_{rand}, a_{missref}, u_{alphas})$$

Where:

- $|a_{sys}|$ is the total of information units present in the synthetic answer, serving as the denominator to normalize the relevance score, ensuring it reflects the proportion of relevant information in relation to the total content volume of the answer.

- The other components of the formula have the same meaning ascribed in the definition of $h_{completeness}$.

The normalization of the sum of similarities ($norm\_s(\cdot)$) is an essential component for both $h_{completeness}$ and $h_{relevance}$. It calculates the contribution of information units added to the synthetic answer in relation to the missing units or the additional information units, adjusted by their respective weights.

$$norm\_s(u_{rand}, a_{missref}, u_{alphas}) = \sum_{i=1}^{|u_{rand}|} \text{sim}(u_{rand}^{(i)}, a_{missref}) \cdot u_{alphas}^{(i)}$$

Where:

- $\text{sim}(u_{rand}^{(i)}, a_{missref})$ represents the similarity between the i-th random information unit and the missing units from the reference answer, maximizing similarity when pertinent.
- $u_{alphas}^{(i)}$ is the weight associated with the i-th random information unit, reflecting its relevance associated with the source of the original information.

These definitions provide the method used to automatically define the completeness and relevance values for a synthetic answer used during the training of regression models.

For each question and answer randomly extracted from the ELI5 and WebGPT datasets, various synthetic versions derived from this original answer are created. This is achieved by varying the value of the $remove\_rate$ parameter, which determines the rate of random information to be removed from the original answer, and the values associated with $inclusion\_quantity$, which determines the amount of information to be randomly added from each additional information source.

By varying the $remove\_rate$ from 0 to 100 in steps of 10, and varying $inclusion\_quantity$ from 0 to 20, with random steps from 0 to 4, a training collection of approximately 100,000 synthetic answers is created from 500 random answers from the ELI5 dataset and 500 random answers from the WebGPT dataset.

The distribution of completeness scores for this set is shown in Figure 6.5 and that of relevance in Figure 6.6. For both cases, a uniform distribution of scores is sought. In the case of completeness, a relatively uniform distribution is observed starting from score 40. However, there is a peak at scores equal to 100 and few scores between 95 and

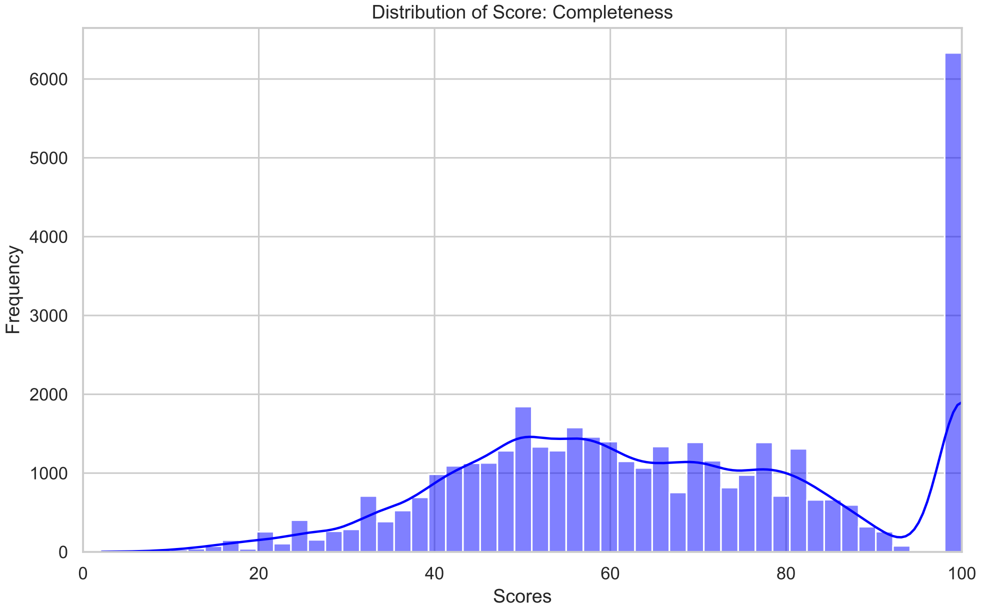


Figure 6.5 – Synthetic completeness scores distributions.
Source: The Author.

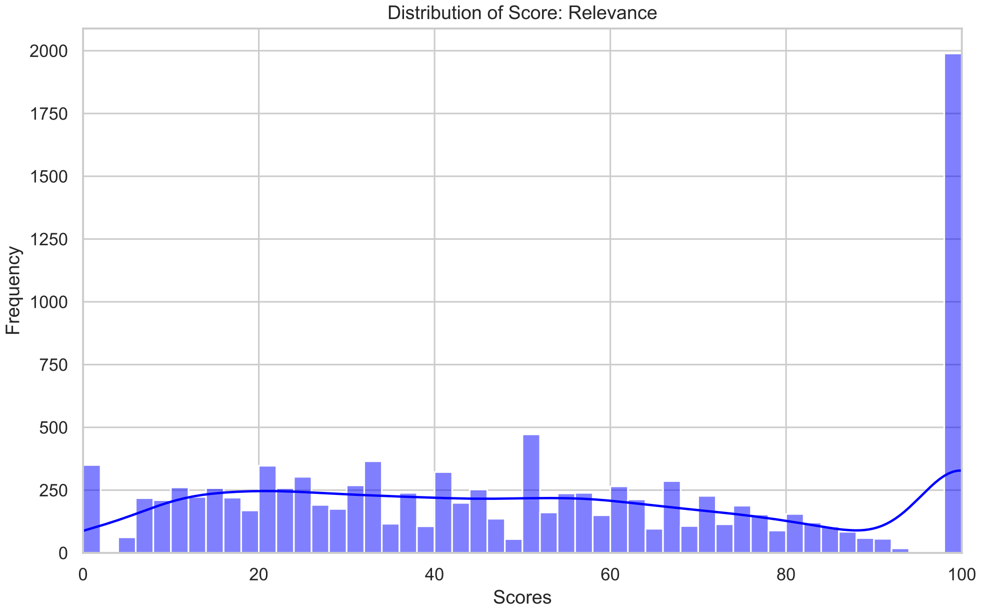


Figure 6.6 – Synthetic relevance scores distributions.
Source: The Author.

99. This is due to the values of the $remove\_rate$ parameter used. When it is equal to 0, completeness is always equal to 100. When it is greater than 0, completeness always turns out to be less than 95. And as in all cases, the original answer is used at least once with the $remove\_rate = 0$, the result is a peak of completeness scores equal to 100. In the case of relevance, a uniform distribution is observed in essentially all scores, except in scores above 95. Again, at least once for each original answer, no extra information was added that could be deemed irrelevant. Thus, there was a peak of relevance scores equal to 100.

This chapter presented the proposed metric models to evaluate the completeness and relevance of answers provided by non-factoid QA systems. Through different types of approaches that combine the use of LLMs with prompt strategies, division of information units, and the training of regression models, differents perspectives were considered for measuring these criteria in evaluating the answers. The next chapter presents the results of the experiments conducted using the dataset proposed in this study. It examines the proposed metrics in conjunction with other baseline metrics to evaluate their effectiveness and compare performance.

7 RESULTS AND ANALYSIS

This chapter aims to present and analyze the results obtained in the experiments of this study, which employed different metric models to assess the criteria of completeness and relevance. It begins with the results and analyses related to the completeness criterion, discussed in Section 7.1, followed by those related to the relevance criterion, presented in Section 7.2. The chapter ends with the final considerations presented in Section 7.3

The results are displayed in tables, where each row corresponds to a tested metric model. The rows are organized into three categories: special metric models (described in Section 4.3), baseline metric models, and metric models proposed by this study. The tables include the following columns:

- **Metric**: name of the evaluated metric model;
- **Ref**: indicates whether the metric model requires a reference answer, marked with *X* when applicable;
- **Spearman**: shows the correlation of the metric model with human evaluations through the Spearman's rank correlation coefficient;
- **Kendall**: similar to Spearman, but using the Kendall rank correlation coefficient;
- **Pearson**: presents the correlation of the metric model with human evaluations by the Pearson correlation coefficient;
- **Accuracy**: based on an accuracy assessment that compares the metric model's preferences answers with human preferences for a unique question, displaying the percentage of times the preferences matched, as described in Section 4.6.2.

The tables also use color highlighting to facilitate comparison between models. A more intense orange color highlights the best result among the models that require a reference answer, while a lighter orange indicates the second best. A stronger blue highlights the best result among the models that do not require a reference answer, and a lighter blue, the second best. Bold values highlight the best results among all evaluated models, except those belonging to the special metrics group, which are not included in the comparison due to being considered of more limited use and not applicable in real world scenarios.

This differentiation of highlights facilitates the identification of the most suitable metric model for each situation. In contexts where sets of reference answers are available, one can choose between models that either require these answers or do not. In situations without reference answers, models that require them are not an option.

Additionally, metric models that utilize generative LLMs, particularly GPT-3.5 and GPT-4 models, are highlighted with a gray background. These models are differentiated by their high number of parameters and the consequent demand for significant computational capacity. To access these GPT models currently, it is necessary to use an API controlled by an external entity, which imposes costs for each request. This characteristic may become the use of these models unfeasible in different contexts. Additionally, although the metric models based in GPT evaluated do not require a reference answer, they were not included in any comparison group (those that use or do not use references). This was done to facilitate the comparison of the models within these groups and because the generative LLMs were considered a separate case, due to their unique characteristics.

For each evaluation criterion, completeness and relevance, three tables are presented according to the previously mentioned characteristics. The first table addresses the complete set of evaluation data, including a more general analysis. This table is unique because it includes the *Accuracy* column, distinguishing it from the others. The calculation of accuracy is possible in this context because it compares the preference between answers generated by humans and by the GPT-4 model with the preferences of human evaluators. Therefore, this metric directly assesses a model's ability to replicate human choices between two distinct options.

The second and third tables focus exclusively on answers created by humans and those generated by the GPT-4 model, respectively. Due to this specialization, the accuracy column is omitted in these cases. This is because accuracy, as defined in the context of this study, requires a direct comparison between a human answer and a GPT-4 answer in relation to the preference of the human evaluator. By focusing only on one type of answer (human or GPT-4), the basis for such a comparison is lost, making the evaluation of accuracy not possible in these tables. Moreover, each of these tables is accompanied by a table with the p-values referring to each correlation coefficient, that is a statistical measure used to determine the significance of the results obtained.

In the interpretation of the results, the Spearman, Kendall, and Pearson correlation coefficients, along with accuracy, reflect the alignment of the metric models with human evaluations in terms of completeness and relevance of the answers. Regarding the difference of each of these criteria:

- **Spearman:** Measures the strength and direction of the association between the rankings of two variables.

- **Kendall:** Focuses on the agreement of ranking pairs, comparing pairs of items to see how many are in the same order.
- **Pearson:** Measures the linear correlation between the metric scores and human evaluations. A high value indicates that as human evaluation increases or decreases, the metric scores adjust proportionally.
- **Accuracy:** Indicates the proportion of times that the metric model's preference coincides with human preference in binary choices (answer generated by a human or the answer generated by GPT-4).

Examples of interpretation:

- If a model's Kendall score is high, but its Spearman score is lower, this may indicate that the model understands local preferences between pairs of items well, but struggles to understand into a global view considering all evaluated instances at once.
- If a model's Kendall score is low, but its Spearman score is higher, this suggests that, although the model may have difficulty consistently identifying the correct preference between specific pairs of answers, it is capable of maintaining a general order of ranking that corresponds to human evaluation.
- If both Kendall and Spearman scores are low and Pearson is high, this indicates that the model has a good capacity to adjust its scores proportionally to variations in human evaluations on a continuous scale, but fails to capture the correct order of preference or agreement of specific pairs.
- If the coefficients are high and accuracy is low, this might indicate that the model is good at replicating the ordering and proportionality of human evaluations considering the set of answers, but fails when choosing the "favorite" answer in direct comparisons between the answer generated by a human and the answer generated by GPT-4.

7.1 Completeness

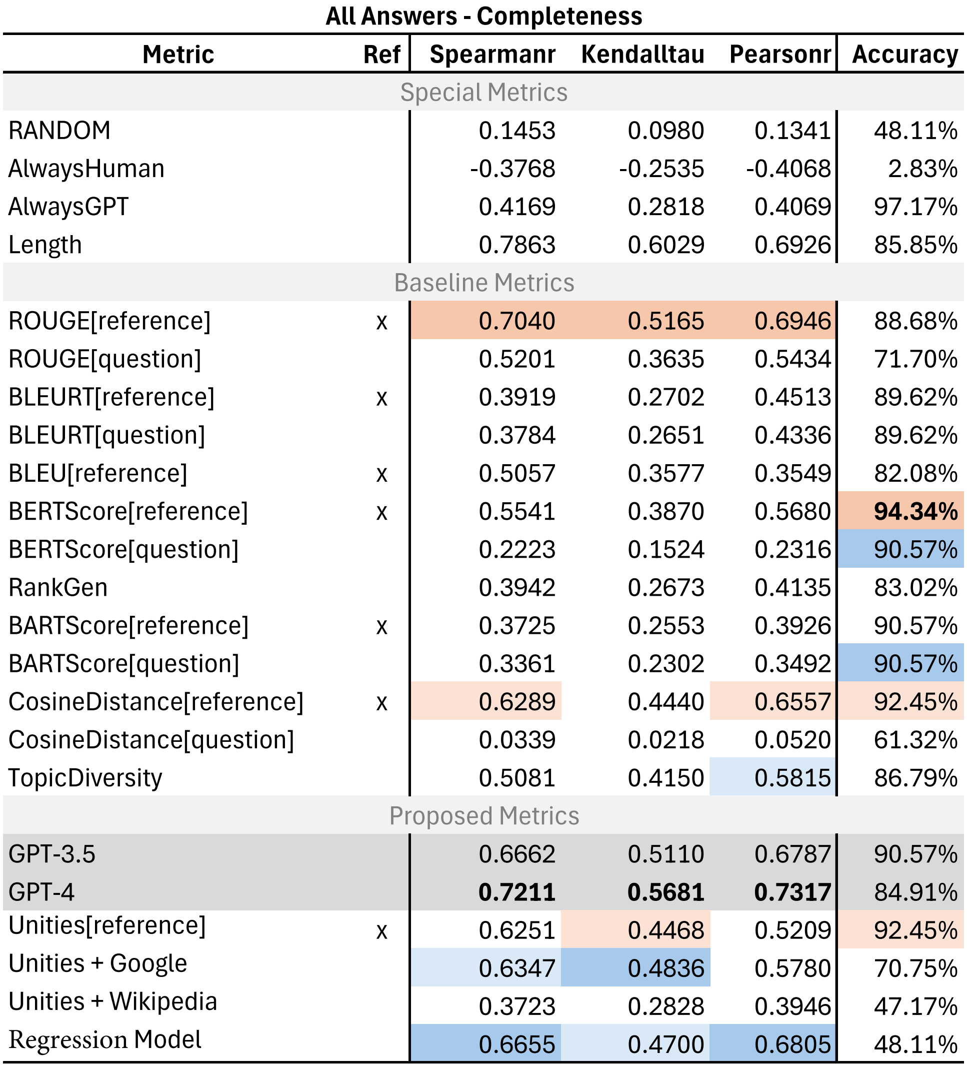
Table 7.1 presents the results obtained by the evaluated metric models according to the completeness criterion for all answers in the evaluation dataset. It was observed that the GPT-4 model became notable by displaying the highest values in the three correlation coefficients: Spearman coefficient (0.7211), Kendall coefficient (0.5681), and

Pearson coefficient (0.7317). These results indicate superior performance in correlation with human evaluations regarding the completeness of the answers. On the other hand, BERTScore[reference] achieved the highest accuracy (94.34%), overcoming GPT-4 by 9.43 percentage points in the task of selecting the preferred answer for a given question according to human evaluators. This notable difference suggests that, although GPT-4 is effective in approximating the human evaluation standard when considering all evaluated answers, BERTScore[reference] demonstrates greater ability in identifying which is the preferred answer to the same question from the perspective of human evaluators.

For metric models that use reference answers, ROUGE[reference] showed the highest values in Spearman, Kendall, and Pearson coefficients (0.7040, 0.5165, and 0.6946, respectively), indicating a strong association with human judgment in the context of completeness. CosineDistance[reference] was the runner-up in Spearman (0.6289) and Pearson (0.6557), and Unities[reference] in Kendall (0.4468). BERTScore[reference] leads in accuracy with 94.34%, 1.89 percentage points above its closest competitor, CosineDistance[reference], with 92.45

Looking at the metrics that do not require reference answers, the Regression Model overcome others in Spearman (0.6655) and Pearson (0.6805), suggesting strong correlations with human completeness judgments. It also achieves the second highest Kendall (0.4700). Unities + Google takes the lead in Kendall (0.4836) and is second in Spearman (0.6347). TopicDiversity is the runner-up in Pearson (0.5815). However, BERTScore[question] and BARTScore[question] share the highest accuracy of 90.57%, indicating that these models are equally proficient in selecting the most complete answer for the same question. These data suggest that, when reference answers are not available, these metrics can still provide evaluations that align well with human judgments on the completeness of the answer.

Table 7.2 presents the p-values for each evaluated metric model for the completeness criterion, considering the set of all answers. In this case, almost all models showed a p-value <0.001, meaning that the reported correlations are statistically significant, and the likelihood that these results occurred by chance is extremely low. This implies there is a statistically significant relationship between the completeness scores assigned by these metric models and the completeness scores as judged by human evaluators. Only the CosineDistance[question] model showed a high p-value, indicating that the relationship between the completeness scores this model assigns using question text with the answer is not statistically significant for the coefficients.



All Answers - Completeness

Metric	Ref	Spearmanr	Kendalltau	Pearsonr	Accuracy
Special Metrics					
RANDOM		0.1453	0.0980	0.1341	48.11%
AlwaysHuman		-0.3768	-0.2535	-0.4068	2.83%
AlwaysGPT		0.4169	0.2818	0.4069	97.17%
Length		0.7863	0.6029	0.6926	85.85%
Baseline Metrics					
ROUGE[reference]	x	0.7040	0.5165	0.6946	88.68%
ROUGE[question]		0.5201	0.3635	0.5434	71.70%
BLEURT[reference]	x	0.3919	0.2702	0.4513	89.62%
BLEURT[question]		0.3784	0.2651	0.4336	89.62%
BLEU[reference]	x	0.5057	0.3577	0.3549	82.08%
BERTScore[reference]	x	0.5541	0.3870	0.5680	**94.34%**
BERTScore[question]		0.2223	0.1524	0.2316	90.57%
RankGen		0.3942	0.2673	0.4135	83.02%
BARTScore[reference]	x	0.3725	0.2553	0.3926	90.57%
BARTScore[question]		0.3361	0.2302	0.3492	90.57%
CosineDistance[reference]	x	0.6289	0.4440	0.6557	92.45%
CosineDistance[question]		0.0339	0.0218	0.0520	61.32%
TopicDiversity		0.5081	0.4150	0.5815	86.79%
Proposed Metrics					
GPT-3.5		0.6662	0.5110	0.6787	90.57%
GPT-4		**0.7211**	**0.5681**	**0.7317**	84.91%
Unities[reference]	x	0.6251	0.4468	0.5209	92.45%
Unities + Google		0.6347	0.4836	0.5780	70.75%
Unities + Wikipedia		0.3723	0.2828	0.3946	47.17%
Regression Model		0.6655	0.4700	0.6805	48.11%

Table 7.1 – All Answers - Completeness.
Source: The Author.

Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	<0.001	<0.001	<0.001
ROUGE[question]	<0.001	<0.001	<0.001
BLEURT[reference]	<0.001	<0.001	<0.001
BLEURT[question]	<0.001	<0.001	<0.001
BLEU[reference]	<0.001	<0.001	<0.001
BERTScore[reference]	<0.001	<0.001	<0.001
BERTScore[question]	0.0011	<0.001	<0.001
RankGen	<0.001	<0.001	<0.001
BARTScore[reference]	<0.001	<0.001	<0.001
BARTScore[question]	<0.001	<0.001	<0.001
CosineDistance[reference]	<0.001	<0.001	<0.001
CosineDistance[question]	0.6233	0.6378	0.4516
TopicDiversity	<0.001	<0.001	<0.001
GPT-3.5	<0.001	<0.001	<0.001
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	<0.001	<0.001	<0.001
Unities + Google	<0.001	<0.001	<0.001
Unities + Wikipedia	<0.001	<0.001	<0.001
Regression Model	<0.001	<0.001	<0.001

Table 7.2 – All Answers - Completeness p-values.
Source: The Author.

7.1.1 Completeness Results with Human Generated Answers

Considering only the answers generated by humans (Table 7.3), GPT-4 maintained the lead in performance in the Pearson coefficient, reaching 0.6846, while the Regression Model highlighted by obtaining the best performance in the Spearman coefficient, with 0.6359, and GPT-3.5 become notable with the best result in the Kendall, marking 0.4751. GPT-4 showed the largest difference compared to the others in the Pearson coefficient, with a margin of 0.0355 percentage points above the second place. This indicates that GPT-4 has a completeness score distribution more aligned with the distribution used by human evaluators.

In the context of models that use references, ROUGE[reference] repeated its highlight, leading in the Spearman and Kendall coefficients (0.6075 and 0.4357, respectively), besides achieving the second position in Pearson, with 0.6281. CosineDistance[reference] presented the second best performance both in Spearman (0.5021) and in Kendall (0.3568), recording the best performance in Pearson, with 0.6340. These results suggest that ROUGE[reference] has a greater ability to identify the most relevant answers within the set, given the considerable performance difference compared to the second place (0.1054 difference in Spearman). On the other hand, the CosineDistance[reference] model is notable in the Pearson coefficient, indicating that its distribution of completeness scores is close to that of human evaluators.

For models that operate without references, the Regression Model demonstrated superiority, achieving the best coefficients: Spearman (0.6359), Kendall (0.4570), and Pearson (0.6491). The Topic Diversity model also became notable, securing the second best mark in Kendall (0.4386) and Pearson (0.6149), while the Unities + Google model obtained the second place in the Spearman coefficient, with 0.5602.

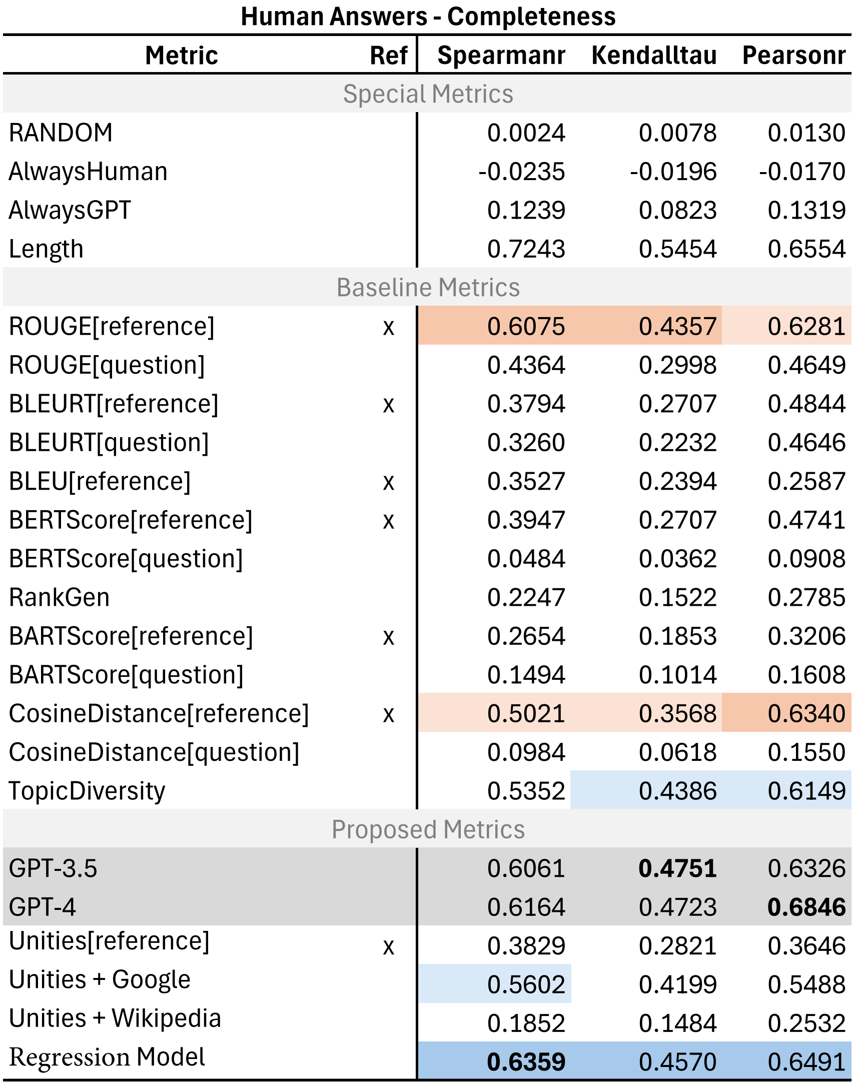
The results also indicate that the highlighted models that do not rely on reference answers outperformed those that use them, with the Regression Model showing results approximately 0.025 points higher than the best reference-based models. This may suggest that the references used had significant differences compared to the answers generated by humans, affecting the performance of models that utilized them. The superior performance of the Regression Model compared to GPT-4 is also noted, being 0.0195 higher in Spearman, which demonstrates the efficiency of the fine-tuned model proposed.

Furthermore, a decrease in performance is observed when considering only answers generated by humans, compared to the performance evaluating all answers. Specifically, GPT-4 experienced a notable decrease of approximately 0.1 points in the three coefficients. There is a decrease in performance in other models as well, suggesting that the evaluation of human answers was more challenging.

Table 7.4 presents the p-values for each evaluated metric model for the completeness criterion, considering the set of answers generated by humans. In this case, the majority of the models showed a p-value <0.001, meaning that the reported correlations are statistically significant. The models BERTScore[question], BARTScore[question], and Unities + Wikipedia showed a high p-value, indicating that the relationship between the completeness scores these models assign and the human evaluations is not statistically significant for the coefficients.

7.1.2 Completeness Results with GPT-4 Generated Answers

When analyzing exclusively the answers generated by the GPT-4 model (Table 7.5), the Regression Model became notable, achieving the best result in all coefficients: Spearman (0.8194), Kendall (0.6227), and Pearson (0.8593). This demonstrates a notable superiority of this model in assigning completeness scores for answers generated by GPT-4, with a significant difference of 0.2106 points compared to the second place in the Spearman coefficient. In the Pearson coefficient, the difference was 0.1413



Human Answers - Completeness

Metric	Ref	Spearmanr	Kendalltau	Pearsonr
Special Metrics				
RANDOM		0.0024	0.0078	0.0130
AlwaysHuman		-0.0235	-0.0196	-0.0170
AlwaysGPT		0.1239	0.0823	0.1319
Length		0.7243	0.5454	0.6554
Baseline Metrics				
ROUGE[reference]	x	0.6075	0.4357	0.6281
ROUGE[question]		0.4364	0.2998	0.4649
BLEURT[reference]	x	0.3794	0.2707	0.4844
BLEURT[question]		0.3260	0.2232	0.4646
BLEU[reference]	x	0.3527	0.2394	0.2587
BERTScore[reference]	x	0.3947	0.2707	0.4741
BERTScore[question]		0.0484	0.0362	0.0908
RankGen		0.2247	0.1522	0.2785
BARTScore[reference]	x	0.2654	0.1853	0.3206
BARTScore[question]		0.1494	0.1014	0.1608
CosineDistance[reference]	x	0.5021	0.3568	0.6340
CosineDistance[question]		0.0984	0.0618	0.1550
TopicDiversity		0.5352	0.4386	0.6149
Proposed Metrics				
GPT-3.5		0.6061	**0.4751**	0.6326
GPT-4		0.6164	0.4723	**0.6846**
Unities[reference]	x	0.3829	0.2821	0.3646
Unities + Google		0.5602	0.4199	0.5488
Unities + Wikipedia		0.1852	0.1484	0.2532
Regression Model		**0.6359**	0.4570	0.6491

Table 7.3 – Human Answers - Completeness.
Source: The Author.

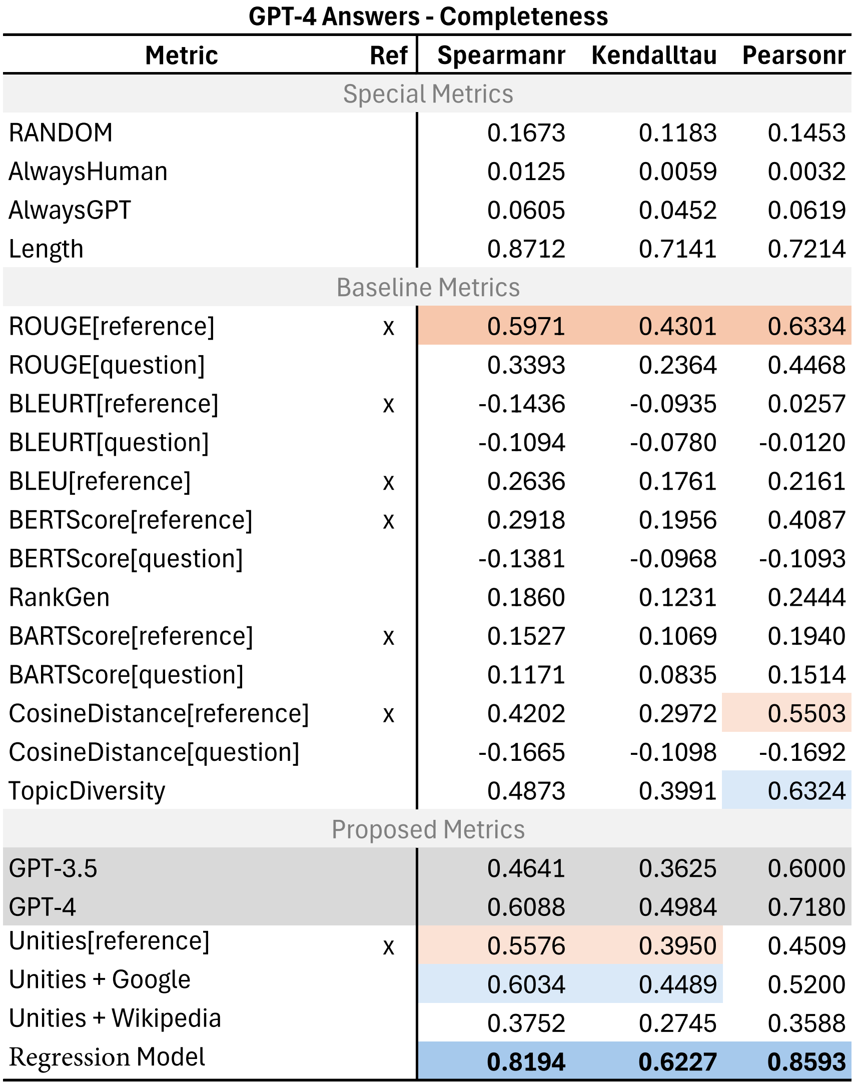
Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	<0.001	<0.001	<0.001
ROUGE[question]	<0.001	<0.001	<0.001
BLEURT[reference]	<0.001	<0.001	<0.001
BLEURT[question]	<0.001	<0.001	<0.001
BLEU[reference]	<0.001	<0.001	0.0074
BERTScore[reference]	<0.001	<0.001	<0.001
BERTScore[question]	0.6224	0.5832	0.3545
RankGen	0.0206	0.0211	0.0038
BARTScore[reference]	0.0060	0.0050	<0.001
BARTScore[question]	0.1265	0.1243	0.0997
CosineDistance[reference]	<0.001	<0.001	<0.001
CosineDistance[question]	0.3155	0.3490	0.1126
TopicDiversity	<0.001	<0.001	<0.001
GPT-3.5	<0.001	<0.001	<0.001
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	<0.001	<0.001	<0.001
Unities + Google	<0.001	<0.001	<0.001
Unities + Wikipedia	0.0574	0.0537	0.0088
Regression Model	<0.001	<0.001	<0.001

Table 7.4 – Human Answers - Completeness p-values.
Source: The Author.

points, also showing a high correlation in the distribution of completeness scores with the evaluations of human annotators.

Considering the models that use references, the ROUGE[reference] model again became notable, achieving the best values in all three coefficients: Spearman (0.5971), Kendall (0.4301), and Pearson (0.6334). The Unities[reference] model was the second best in Spearman (0.5576) and Kendall (0.3950), while the CosineDistance[reference] model obtained the second best result in Pearson (0.5503). These results suggest a superiority of the ROUGE[reference] model among those using a reference to evaluate completeness, as it performed best in most comparative tests.

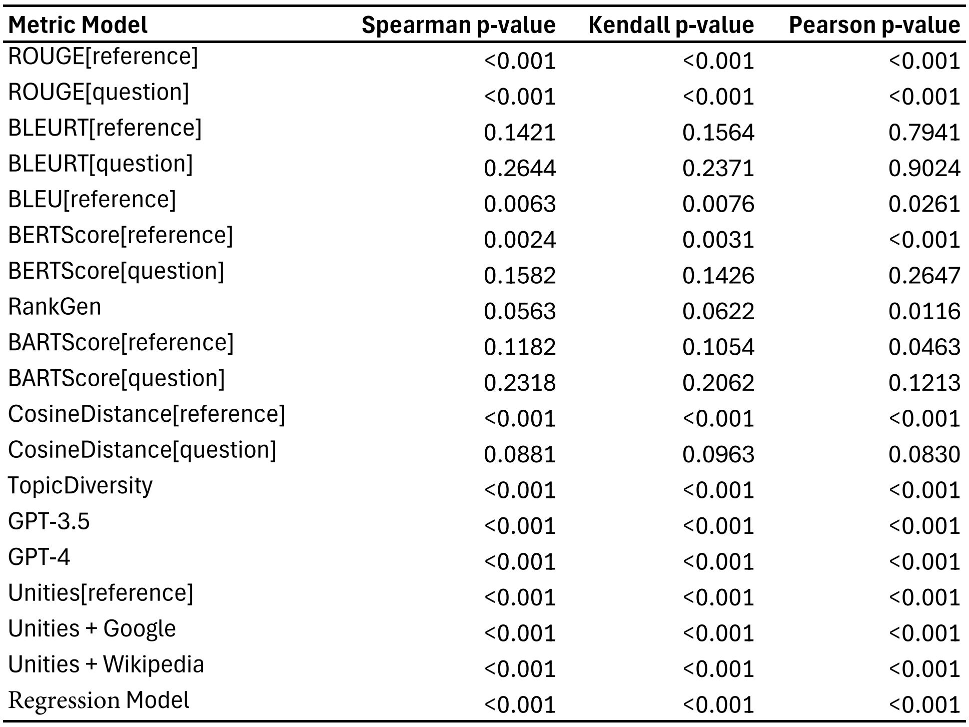
For models that operate without the need for a reference, the Regression Model again highlighted with the best coefficients, as already mentioned. The Unities + Wikipedia model also showed promising results, achieving the second best performance in Spearman (0.6034) and Kendall (0.4489). CosineDistance[reference] recorded the second best in Pearson (0.6324), standing out among the evaluated models. Once again, the models that do not use reference answers outperformed those that do, indicating the viability of obtaining competitive results without the use of this resource. Moreover, when focusing on this set of answers generated by GPT-4, these models significantly outperformed the performance of GPT-4, especially the Regression Model, which achieved considerably superior results.



GPT-4 Answers - Completeness

Metric	Ref	Spearmanr	Kendalltau	Pearsonr
Special Metrics				
RANDOM		0.1673	0.1183	0.1453
AlwaysHuman		0.0125	0.0059	0.0032
AlwaysGPT		0.0605	0.0452	0.0619
Length		0.8712	0.7141	0.7214
Baseline Metrics				
ROUGE[reference]	x	0.5971	0.4301	0.6334
ROUGE[question]		0.3393	0.2364	0.4468
BLEURT[reference]	x	-0.1436	-0.0935	0.0257
BLEURT[question]		-0.1094	-0.0780	-0.0120
BLEU[reference]	x	0.2636	0.1761	0.2161
BERTScore[reference]	x	0.2918	0.1956	0.4087
BERTScore[question]		-0.1381	-0.0968	-0.1093
RankGen		0.1860	0.1231	0.2444
BARTScore[reference]	x	0.1527	0.1069	0.1940
BARTScore[question]		0.1171	0.0835	0.1514
CosineDistance[reference]	x	0.4202	0.2972	0.5503
CosineDistance[question]		-0.1665	-0.1098	-0.1692
TopicDiversity		0.4873	0.3991	0.6324
Proposed Metrics				
GPT-3.5		0.4641	0.3625	0.6000
GPT-4		0.6088	0.4984	0.7180
Unities[reference]	x	0.5576	0.3950	0.4509
Unities + Google		0.6034	0.4489	0.5200
Unities + Wikipedia		0.3752	0.2745	0.3588
Regression Model		**0.8194**	**0.6227**	**0.8593**

Table 7.5 – GPT-4 Answers - Completeness.
Source: The Author.



Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	<0.001	<0.001	<0.001
ROUGE[question]	<0.001	<0.001	<0.001
BLEURT[reference]	0.1421	0.1564	0.7941
BLEURT[question]	0.2644	0.2371	0.9024
BLEU[reference]	0.0063	0.0076	0.0261
BERTScore[reference]	0.0024	0.0031	<0.001
BERTScore[question]	0.1582	0.1426	0.2647
RankGen	0.0563	0.0622	0.0116
BARTScore[reference]	0.1182	0.1054	0.0463
BARTScore[question]	0.2318	0.2062	0.1213
CosineDistance[reference]	<0.001	<0.001	<0.001
CosineDistance[question]	0.0881	0.0963	0.0830
TopicDiversity	<0.001	<0.001	<0.001
GPT-3.5	<0.001	<0.001	<0.001
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	<0.001	<0.001	<0.001
Unities + Google	<0.001	<0.001	<0.001
Unities + Wikipedia	<0.001	<0.001	<0.001
Regression Model	<0.001	<0.001	<0.001

Table 7.6 – GPT-4 Answers - Completeness p-values.
Source: The Author.

Table 7.6 presents the p-values for each metric model evaluated for the completeness criterion considering the set of answers generated by the GPT-4 model. In this case, most models showed a p-value <0.001, indicating that the reported correlations are statistically significant. The BLEURT[reference], BLEURT[question], BERTScore[question], RankGen, BARTScore[reference], BARTScore[question], and CosineDistance[question] models showed a high p-value, indicating that the relationship between the completeness scores assigned by this model and the human evaluations is not statistically significant for the coefficients.

It is observed that, when considering all the answers, including those generated by humans and by GPT-4, the models using a reference answer achieve a higher correlation score with human annotations than those that do not use this feature. However, when examining the performance of the models individually for the sets of answers generated by humans or by GPT-4, the models without reference answers prove to be considerably superior in evaluating completeness. This can be explained by the functioning of the correlation coefficients. Spearman and Kendall assess the preference order of all answers, suggesting that, when using the total set of answers, the models with reference can produce a preference list more similar to that of human evaluators. On the other hand, when focusing on individual sets, the models without reference demonstrate to create preference lists more aligned with those of human evaluators.

7.1.3 Discussions About Completeness Results

The models that became notable for their performance include GPT-4, the Regression Model, ROUGE[reference], and CosineDistance[reference]. In the case of GPT-4 with the proposed prompt approach, thanks to its characteristics related to the vast number of parameters that enable efficient information storage, it is presumed to have a remarkable capacity to understand the external world. This contributes to better understand regarding which are all the needed relevant information to a specific question. This capacity is a possible explanation for its performance in completeness tests, where it maintained competitive scores across different situations. Notably, GPT-4 is one of the few models that achieve a high score across the three correlation coefficients and the accuracy criterion simultaneously. This suggests that GPT-4 shows remarkable performance both in ordering the most complete answers from a set and in pair-wise comparison, while achieving a score distribution similar to that of human annotators and being capable of determining the preferred answer, in terms of completeness, between one generated by humans and one generated by GPT-4 for the same question.

The Regression Model became notable for its superior score compared to other metrics that do not use reference answers, even overcoming the performance of GPT-4 in different scenarios. This demonstrates the effectiveness of the fine-tuning approach proposed with a synthetic dataset, designed to simulate conditions of more or less complete answers. This model can be considered less costly in terms of application compared to models like GPT-4, especially because it is considerably smaller, in terms of computational resources required and do not require a paid API access, in addition to does not depend on reference answers, facilitating its implementation in different contexts where these resources are not available.

Another model that showed competitive performance across all criteria was ROUGE[reference]. Although this model depends on a reference answer, it has the ability to identify which relevant information is missing in the evaluated answer, possibly through measuring the overlap of lexical units between the text of the evaluated answer and the reference text.

CosineDistance[reference] is another metric model that became notable across all criteria. Using a reference answer, this model seeks to represent all relevant information from the reference answer in a vector of embeddings, employing the BERT model and calculating the average across tokens. Similarly, it determines the embedding vector of the evaluated answer and calculates the distance to the vector of the reference answer. Greater distances indicate that relevant information may be missing in the evaluated answer, which consequently reduces the completeness score.

The TopicDiversity model also became notable across all criteria. Unlike other evaluated models, TopicDiversity is distinguished by not using any external resources, not even the question text. A possible explanation for its high performance is that answers with lower topic diversity tend to be more complete than those with high diversity, once these answers have information focused on a single topic, allowing for greater detail.

The Unities + Google model also became notable, presenting competitive scores in various situations. This proposed model is mainly differentiated by not requiring reference answers to function, needing only documents relevant to the question, such as Google pages. This characteristic may simplify its application in different contexts. These results indicate that the approach replicating the recall metric to determine completeness shows promise. On the other hand, the Unities + Wikipedia model, despite operating in a manner similar to the Unities + Google, which showed competitive scores, performed poorly. This model utilizes Wikipedia articles as reference documents. From this, one can infer that the quality or relevance of the reference documents used in the Unities approach has a significant impact on the results achieved.

Metric models such as BLEURT[reference], BLEURT[question], BLEU[reference], BERTScore[reference], RankGen, and BARTScore[reference] exhibited average performances when compared to other models. These metrics, used for the general evaluation of text, indicate that although they may reflect the completeness criterion to a certain extent, they are not specialized enough in this aspect to compete with models focused on completeness, such as GPT-4 and the Regression Model.

Metric models like CosineDistance[question], BERTScore[question], and BARTScore[question] showed inferior performance compared to other metrics. These models used the question as a reference, instead of an answer. This suggests that these metrics require reference answers to work effectively, and not just the text of the question. In contrast, ROUGE[question], which also uses the question as a reference, achieved scores closer to the models that employ a reference answer. This indicates that, among the general text evaluation metrics, ROUGE[question] may be more suitable for situations where a reference answer is not available.

Regarding the accuracy metric, used to evaluate the metric models, the special metric AlwaysGPT revealed that annotators generally show a preference for the answers produced by GPT-4. Therefore, metrics that tend to assign higher scores to answers generated by GPT-4 have an advantage in this specific criterion. This characteristic was particularly observed in the BERTScore, BARTScore, CosineDistance[reference], and Unities[reference] models, which became notable for presenting high accuracy. Thus, when metric models show low correlation coefficients but high accuracy, it can be inferred that they have a tendency to assign a higher score to answers produced by the GPT-4 model instead of considering completeness characteristics.

Another trend observed with special metrics is related to the length of the answer. Intuitively, the longer the answer, the more complete it may seems. Therefore, noticing high scores in the special Length metric, we can conclude that evaluators tend to assign higher scores to longer answers. This behavior is expected, as for an answer to be considered complete by the evaluator, it must contain a significant amount of relevant information. Furthermore, from the accuracy of the Length metric, it is observed that human evaluators also tend to prefer longer answers to the same question.

As for tests with different sets of answers, both those generated by humans and those produced by GPT-4, there was a significant change in the performance of the evaluated metrics. This signals a sensitivity of the models to the text style of the answers. In general, metric models performed better when evaluating the entire set of answers, including those produced by humans and by GPT-4. It was in the set of answers generated exclusively by GPT-4 that most models encountered more difficulties. The exception was the Regression Model, which showed high scores in correlation coefficients.

7.2 Relevance

Table 7.7 presents the results obtained by the evaluated metric models according to the relevance criterion for all answers in the evaluation dataset. It is observed that the GPT-4 model consistently showed the best performance in the three correlation coefficients: Spearman's coefficient (0.5927), Kendall's coefficient (0.4623), and Pearson's coefficient (0.6681). On the other hand, Unities[reference] achieved the highest accuracy (92.45%), overcoming GPT-4 by 1.88 percentage points in the criterion of preferring the same answer for a given question according to human evaluators, considering the relevance criterion. This small difference with the first place also demonstrates that GPT-4 has the ability to identify which is the preferred answer for the same question from the perspective of human evaluators.

For the metric models that use reference answers, BERTScore[reference] presented the highest values in the Spearman, Kendall, and Pearson correlation coefficients (0.5374, 0.3818, and 0.5251, respectively), indicating a considerable association with human judgment in the context of relevance. Unities[reference] is the runner-up in the three coefficients: Spearman's coefficient (0.5140), Kendall's coefficient (0.3749), and Pearson's coefficient (0.5063). Moreover, Unities[reference] leads in accuracy, with 92.45%. In second place in accuracy comes the BLEURT[reference] model with 82.08%.

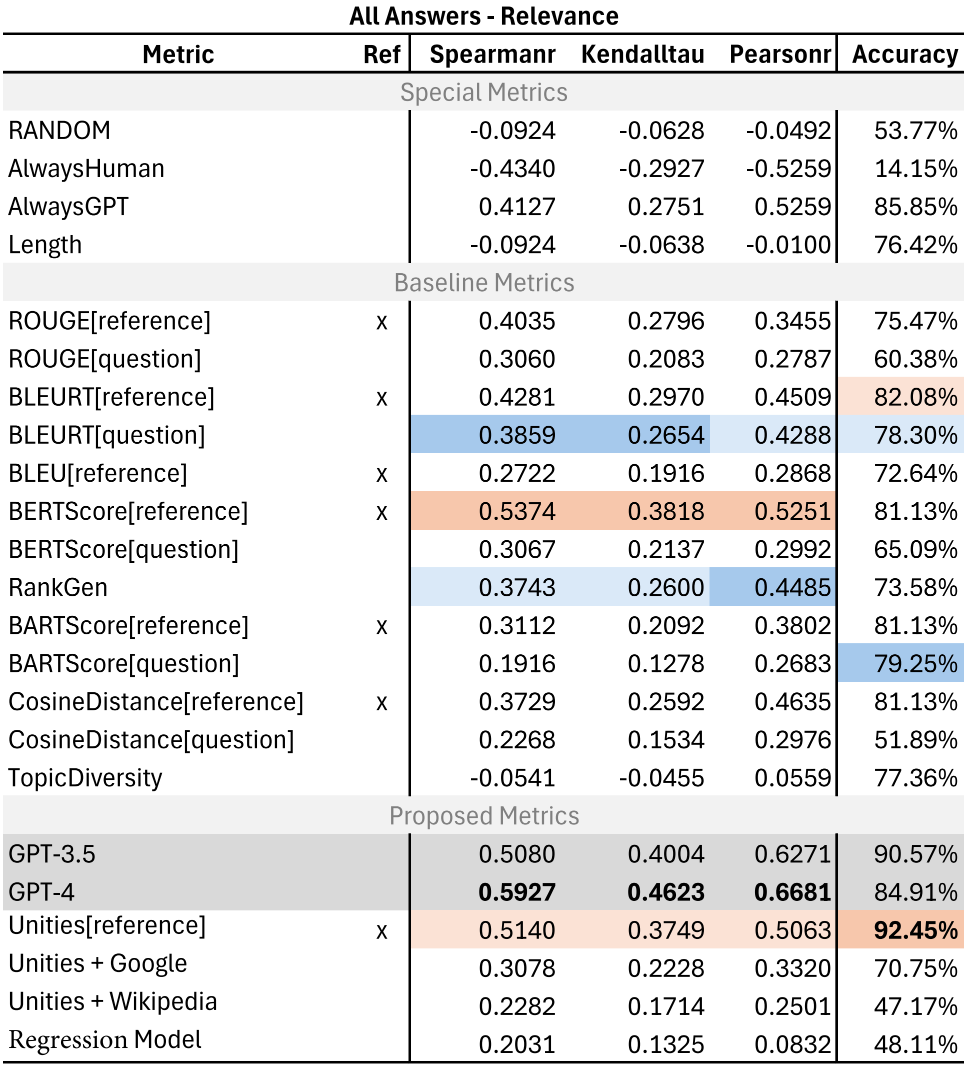
Looking at the metrics that do not require reference answers, BLEURT[question] presents the highest Spearman's coefficient (0.3859), the highest Kendall's coefficient (0.2654), and the second highest Pearson's coefficient (0.4288). Meanwhile, the RankGen model showed the second best result in the Spearman's coefficient (0.3743) and Kendall's coefficient (0.3859), and the best result in the Pearson's coefficient (0.4485). These results present a significant difference of 0.1518 (Spearman) lower compared to the best metric that uses a reference answer. This indicates a difficulty in evaluating relevance without a reference answer.

Table 7.8 presents the p-values for each evaluated metric model according to the relevance criterion considering the set of all answers. In this case, the vast majority of the models showed a p-value <0.001, which means that the reported correlations are statistically significant. The BARTScore[question] and TopicDiversity models presented a high p-value, indicating that the relationship between the completeness scores assigned by these models and human evaluations is not statistically significant for the coefficients. Also, the Regression Model showed a high p-value for Pearson's coefficient.

7.2.1 Relevance Results with Human Generated Answers

Considering only the answers generated by humans (Table 7.9), GPT-4 maintained the lead in the three correlation coefficients: Spearman (0.5382), Kendall (0.4063), and Pearson (0.5809). This performance is slightly below the performance with the test set with all answers, with 0.0545 (Spearman), 0.0560 (Kendall), and 0.0872 (Pearson) lower, indicating a slight increase in difficulty with answers exclusively created by humans.

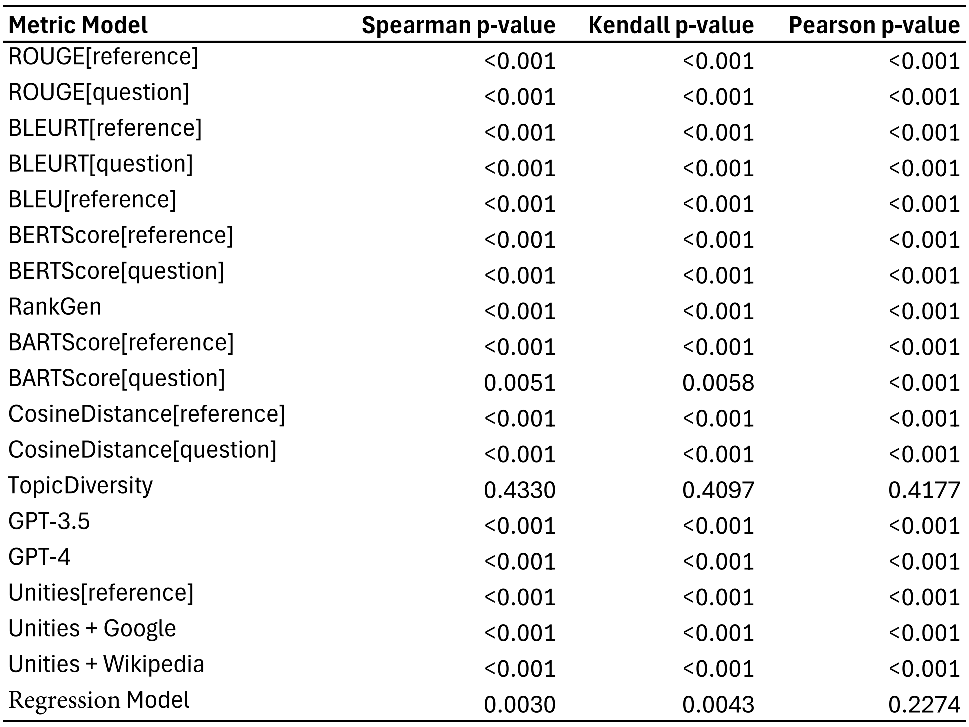
In the context of models that use references, BERTScore[reference] again achieved the best results, leading in the three coefficients: Spearman (0.4996), Kendall (0.3480), and Pearson (0.5008). This time, the ROUGE[reference] model achieved the second best result in the Spearman (0.3789) and Kendall (0.2674) coefficients, while the



All Answers - Relevance

Metric	Ref	Spearmanr	Kendalltau	Pearsonr	Accuracy
Special Metrics					
RANDOM		-0.0924	-0.0628	-0.0492	53.77%
AlwaysHuman		-0.4340	-0.2927	-0.5259	14.15%
AlwaysGPT		0.4127	0.2751	0.5259	85.85%
Length		-0.0924	-0.0638	-0.0100	76.42%
Baseline Metrics					
ROUGE[reference]	x	0.4035	0.2796	0.3455	75.47%
ROUGE[question]		0.3060	0.2083	0.2787	60.38%
BLEURT[reference]	x	0.4281	0.2970	0.4509	82.08%
BLEURT[question]		0.3859	0.2654	0.4288	78.30%
BLEU[reference]	x	0.2722	0.1916	0.2868	72.64%
BERTScore[reference]	x	0.5374	0.3818	0.5251	81.13%
BERTScore[question]		0.3067	0.2137	0.2992	65.09%
RankGen		0.3743	0.2600	0.4485	73.58%
BARTScore[reference]	x	0.3112	0.2092	0.3802	81.13%
BARTScore[question]		0.1916	0.1278	0.2683	79.25%
CosineDistance[reference]	x	0.3729	0.2592	0.4635	81.13%
CosineDistance[question]		0.2268	0.1534	0.2976	51.89%
TopicDiversity		-0.0541	-0.0455	0.0559	77.36%
Proposed Metrics					
GPT-3.5		0.5080	0.4004	0.6271	90.57%
GPT-4		**0.5927**	**0.4623**	**0.6681**	84.91%
Unities[reference]	x	0.5140	0.3749	0.5063	**92.45%**
Unities + Google		0.3078	0.2228	0.3320	70.75%
Unities + Wikipedia		0.2282	0.1714	0.2501	47.17%
Regression Model		0.2031	0.1325	0.0832	48.11%

Table 7.7 – All Answers - Relevance.
Source: The Author.



Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	<0.001	<0.001	<0.001
ROUGE[question]	<0.001	<0.001	<0.001
BLEURT[reference]	<0.001	<0.001	<0.001
BLEURT[question]	<0.001	<0.001	<0.001
BLEU[reference]	<0.001	<0.001	<0.001
BERTScore[reference]	<0.001	<0.001	<0.001
BERTScore[question]	<0.001	<0.001	<0.001
RankGen	<0.001	<0.001	<0.001
BARTScore[reference]	<0.001	<0.001	<0.001
BARTScore[question]	0.0051	0.0058	<0.001
CosineDistance[reference]	<0.001	<0.001	<0.001
CosineDistance[question]	<0.001	<0.001	<0.001
TopicDiversity	0.4330	0.4097	0.4177
GPT-3.5	<0.001	<0.001	<0.001
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	<0.001	<0.001	<0.001
Unities + Google	<0.001	<0.001	<0.001
Unities + Wikipedia	<0.001	<0.001	<0.001
Regression Model	0.0030	0.0043	0.2274

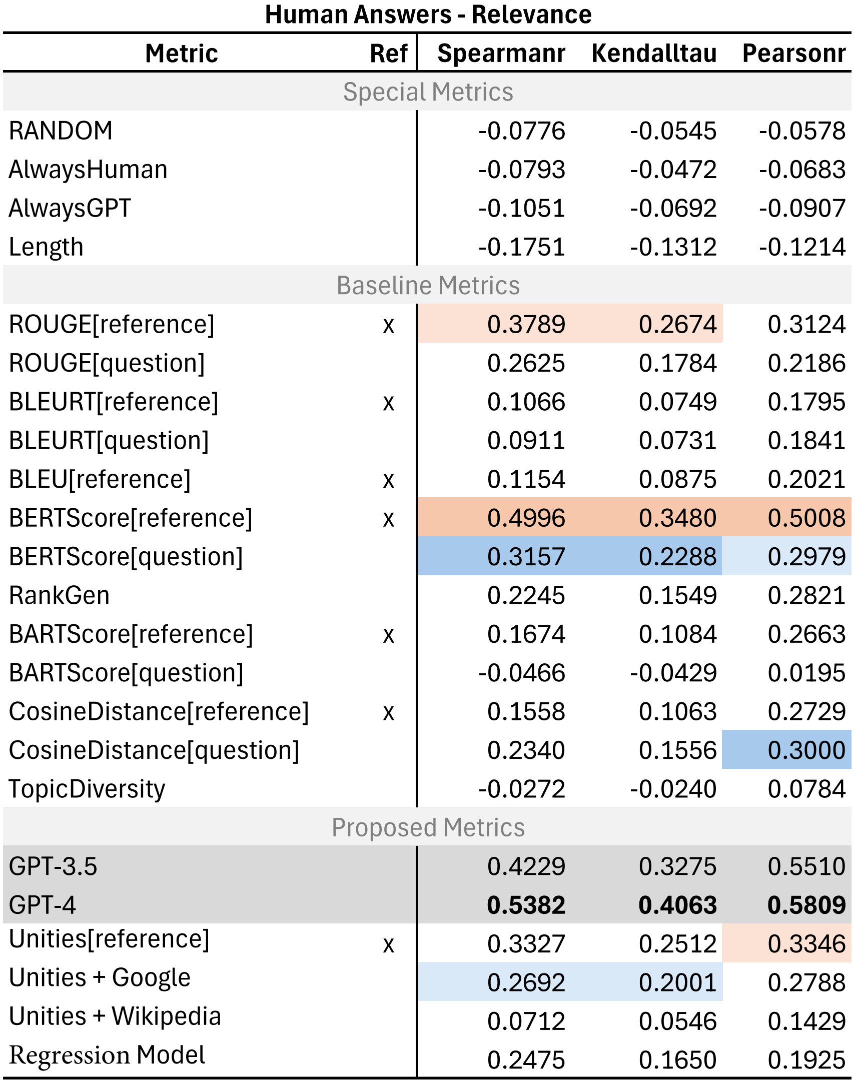
Table 7.8 – All Answers - Relevance p-values.
Source: The Author.

Unities[reference] model achieved the second best result in Pearson (0.3346). These results show the superiority of the BERTScore[reference] model also in the set of human answers.

For the models that operate without references, the BERTScore[question] model demonstrated superiority in this set, achieving the best Spearman (0.3157), Kendall (0.2288), and the second best Pearson (0.2979) coefficients. CosineDistance[question] secured the best mark in Pearson (0.3000), while the Unities + Google model obtained the second place in the Spearman (0.2692) and Kendall (0.2001) coefficients. Again, these results demonstrate a low performance of the metric models that do not use a reference answer in comparison to those that do, indicating a trend of the necessity for a reference for the relevance criterion.

A decline in performance is also observed in the relevance criterion when considering only answers generated by humans, compared to the performance evaluating all answers test set, as there is a decrease in model performance, suggesting that the evaluation of human answers was more challenging.

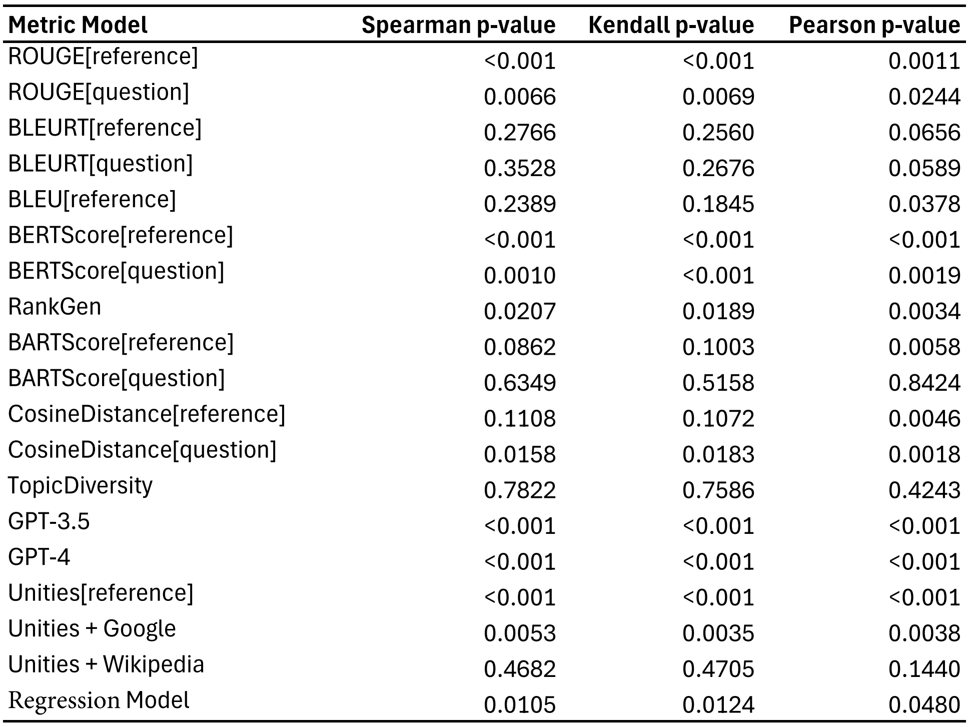
Table 7.10 presents the p-values for each evaluated metric model according to the relevance criterion considering the set of answers generated by humans. In this case, some models presented a p-value < 0.05, which means that the reported correlations are statistically significant for this study. Many of the other models showed a high p-value,



Human Answers - Relevance

Metric	Ref	Spearmanr	Kendalltau	Pearsonr
Special Metrics				
RANDOM		-0.0776	-0.0545	-0.0578
AlwaysHuman		-0.0793	-0.0472	-0.0683
AlwaysGPT		-0.1051	-0.0692	-0.0907
Length		-0.1751	-0.1312	-0.1214
Baseline Metrics				
ROUGE[reference]	x	0.3789	0.2674	0.3124
ROUGE[question]		0.2625	0.1784	0.2186
BLEURT[reference]	x	0.1066	0.0749	0.1795
BLEURT[question]		0.0911	0.0731	0.1841
BLEU[reference]	x	0.1154	0.0875	0.2021
BERTScore[reference]	x	0.4996	0.3480	0.5008
BERTScore[question]		0.3157	0.2288	0.2979
RankGen		0.2245	0.1549	0.2821
BARTScore[reference]	x	0.1674	0.1084	0.2663
BARTScore[question]		-0.0466	-0.0429	0.0195
CosineDistance[reference]	x	0.1558	0.1063	0.2729
CosineDistance[question]		0.2340	0.1556	0.3000
TopicDiversity		-0.0272	-0.0240	0.0784
Proposed Metrics				
GPT-3.5		0.4229	0.3275	0.5510
GPT-4		**0.5382**	**0.4063**	**0.5809**
Unities[reference]	x	0.3327	0.2512	0.3346
Unities + Google		0.2692	0.2001	0.2788
Unities + Wikipedia		0.0712	0.0546	0.1429
Regression Model		0.2475	0.1650	0.1925

Table 7.9 – Human Answers - Relevance.
Source: The Author.



Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	<0.001	<0.001	0.0011
ROUGE[question]	0.0066	0.0069	0.0244
BLEURT[reference]	0.2766	0.2560	0.0656
BLEURT[question]	0.3528	0.2676	0.0589
BLEU[reference]	0.2389	0.1845	0.0378
BERTScore[reference]	<0.001	<0.001	<0.001
BERTScore[question]	0.0010	<0.001	0.0019
RankGen	0.0207	0.0189	0.0034
BARTScore[reference]	0.0862	0.1003	0.0058
BARTScore[question]	0.6349	0.5158	0.8424
CosineDistance[reference]	0.1108	0.1072	0.0046
CosineDistance[question]	0.0158	0.0183	0.0018
TopicDiversity	0.7822	0.7586	0.4243
GPT-3.5	<0.001	<0.001	<0.001
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	<0.001	<0.001	<0.001
Unities + Google	0.0053	0.0035	0.0038
Unities + Wikipedia	0.4682	0.4705	0.1440
Regression Model	0.0105	0.0124	0.0480

Table 7.10 – Human Answers - Relevance p-values.

Source: The Author.

indicating that the relationship between the completeness scores assigned by these models and human evaluations is not statistically significant for the coefficients.

7.2.2 Relevance Results with GPT-4 Generated Answers

When analyzing exclusively the answers generated by the GPT-4 model (Table 7.11), the GPT-4 model remained superior across all coefficients: Spearman (0.3670), Kendall (0.2913), and Pearson (0.3710). Furthermore, the GPT-4 model showed a significant difference of 0.1443 points compared to the second place in the Spearman coefficient. In the Pearson coefficient, the difference was 0.1573 points, also indicating a high correlation in the distribution of relevance scores with the evaluations from human annotators.

Considering the models that use references, BERTScore[reference] again achieved the best results, leading in all three coefficients: Spearman (0.2227), Kendall (0.1607), and Pearson (0.1070). The second position is divided between the Unities[reference] model, with the second highest Spearman (0.1362), the ROUGE[reference] model, with the second highest Kendall (0.0930), and the BLEURT[reference] model, with the second highest Pearson (0.0789). These results suggest a superiority of the ROUGE[reference] model among those that use references to evaluate relevance, since it achieved the best performance in most comparative tests.

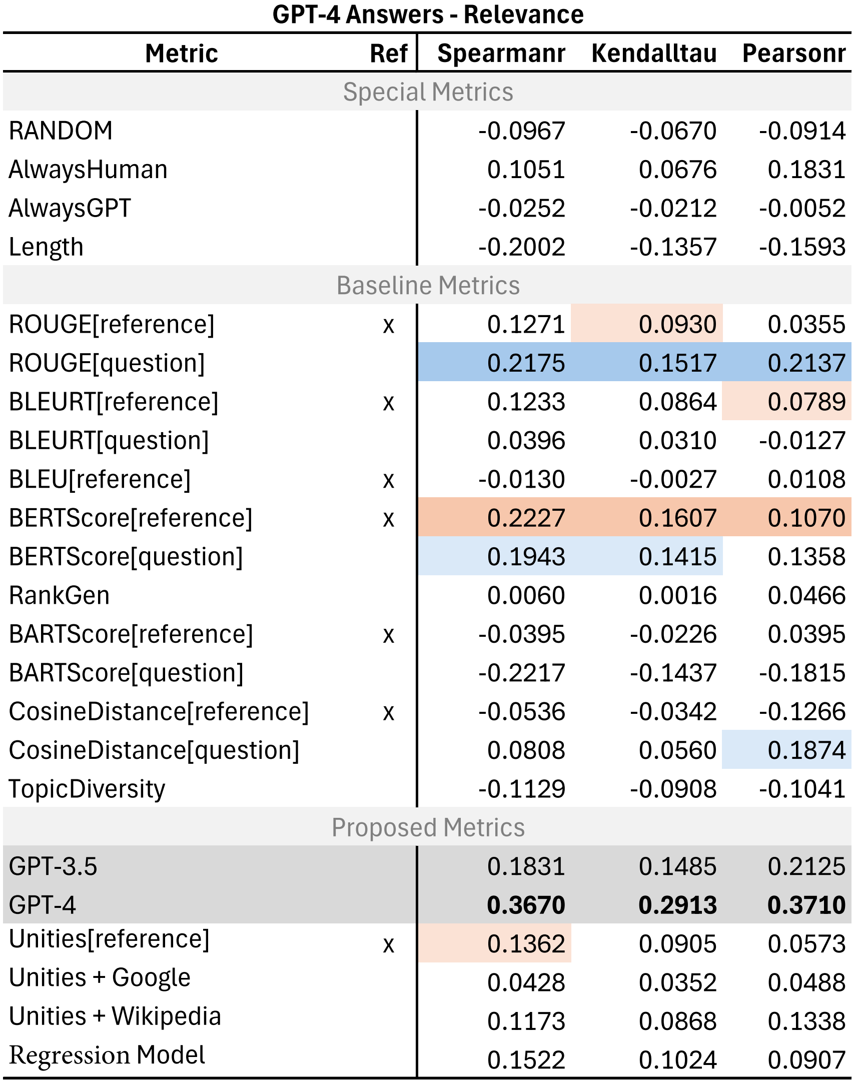
Among the models that operate without the need for references, ROUGE is notable by presenting the best coefficients: Spearman (0.2175), Kendall (0.1517), and Pearson (0.2137). BERTScore achieved the second best performance in Spearman (0.1943) and Kendall (0.1415), while CosineDistance recorded the second best result in Pearson (0.1874). For the set of answers generated by GPT-4, the models without references achieved results similar to those of the models that require references, with a marginal difference of 0.0052 in the Spearman coefficient. This slight difference can be attributed to the difficulty of the models in evaluating the answers produced by GPT-4. In other words, the level of difficulty was considerably raised, such that the metric models were unable to stand out over the others, even those that used reference answers.

In the evaluation of metric models using the set of answers generated by GPT-4, a low score in correlation coefficients is noted. This observation is confirmed by the average scores among the result tables. Considering only the answers generated by GPT-4, the average Spearman coefficient for the evaluated models is 0.2259 lower than the average of the models considering all answers. Furthermore, a generally low performance is observed in the model scores when evaluating the set of answers generated by GPT-4, with an average of 0.1664 for the Spearman coefficient.

Table 7.12 presents the p-values for each metric model evaluated for the relevance criterion considering the set of answers generated by the GPT-4 model. In this case, only some models showed a p-value < 0.05 in some coefficients, which means that the reported correlations, for the coefficient in question, are statistically significant for this study. A good portion of the other models showed a high p-value, indicating that the relationship between the completeness scores that this model assigns and the human evaluations is not statistically significant for the coefficients.

7.2.3 Discussions About Relevance results

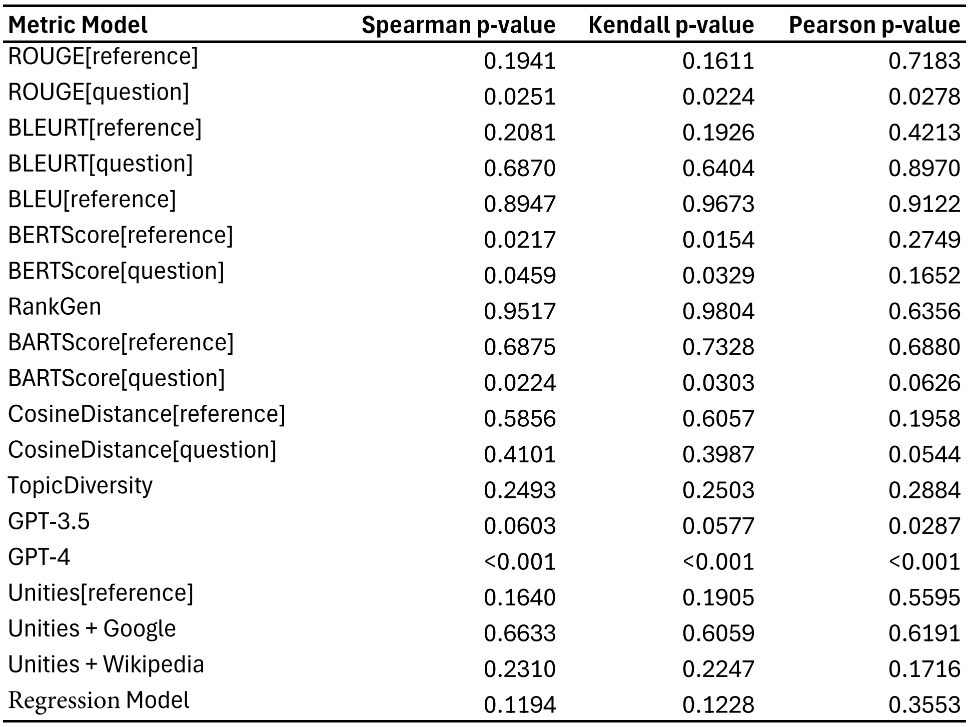
The models that became notable in the experiments with relevance were GPT-4, GPT-3.5, BERTScore[reference], Unities[reference], BLEURT[reference], and ROUGE[reference]. Notably, disregarding the GPT models, all highlighted models use reference-based. This might indicate a trend towards the necessity for reference-based metric models to determine the relevance score. In comparison, models that do not use a reference answer generally demonstrated low performance in assigning relevance scores.



GPT-4 Answers - Relevance

Metric	Ref	Spearmanr	Kendalltau	Pearsonr
Special Metrics				
RANDOM		-0.0967	-0.0670	-0.0914
AlwaysHuman		0.1051	0.0676	0.1831
AlwaysGPT		-0.0252	-0.0212	-0.0052
Length		-0.2002	-0.1357	-0.1593
Baseline Metrics				
ROUGE[reference]	x	0.1271	0.0930	0.0355
ROUGE[question]		0.2175	0.1517	0.2137
BLEURT[reference]	x	0.1233	0.0864	0.0789
BLEURT[question]		0.0396	0.0310	-0.0127
BLEU[reference]	x	-0.0130	-0.0027	0.0108
BERTScore[reference]	x	0.2227	0.1607	0.1070
BERTScore[question]		0.1943	0.1415	0.1358
RankGen		0.0060	0.0016	0.0466
BARTScore[reference]	x	-0.0395	-0.0226	0.0395
BARTScore[question]		-0.2217	-0.1437	-0.1815
CosineDistance[reference]	x	-0.0536	-0.0342	-0.1266
CosineDistance[question]		0.0808	0.0560	0.1874
TopicDiversity		-0.1129	-0.0908	-0.1041
Proposed Metrics				
GPT-3.5		0.1831	0.1485	0.2125
GPT-4		**0.3670**	**0.2913**	**0.3710**
Unities[reference]	x	0.1362	0.0905	0.0573
Unities + Google		0.0428	0.0352	0.0488
Unities + Wikipedia		0.1173	0.0868	0.1338
Regression Model		0.1522	0.1024	0.0907

Table 7.11 – GPT-4 Answers - Relevance.
Source: The Author.



Metric Model	Spearman p-value	Kendall p-value	Pearson p-value
ROUGE[reference]	0.1941	0.1611	0.7183
ROUGE[question]	0.0251	0.0224	0.0278
BLEURT[reference]	0.2081	0.1926	0.4213
BLEURT[question]	0.6870	0.6404	0.8970
BLEU[reference]	0.8947	0.9673	0.9122
BERTScore[reference]	0.0217	0.0154	0.2749
BERTScore[question]	0.0459	0.0329	0.1652
RankGen	0.9517	0.9804	0.6356
BARTScore[reference]	0.6875	0.7328	0.6880
BARTScore[question]	0.0224	0.0303	0.0626
CosineDistance[reference]	0.5856	0.6057	0.1958
CosineDistance[question]	0.4101	0.3987	0.0544
TopicDiversity	0.2493	0.2503	0.2884
GPT-3.5	0.0603	0.0577	0.0287
GPT-4	<0.001	<0.001	<0.001
Unities[reference]	0.1640	0.1905	0.5595
Unities + Google	0.6633	0.6059	0.6191
Unities + Wikipedia	0.2310	0.2247	0.1716
Regression Model	0.1194	0.1228	0.3553

Table 7.12 – GPT Answers - Relevance p-values.
Source: The Author.

This observation mainly considers the Spearman and Kendall coefficients, which represent ranking correlation measures. On the other hand, considering the Pearson coefficient, which measures the linear correlation between two variables, the metric scores of both groups were closer, indicating a correlation of relevance score distributions that are more aligned. This suggests that although reference-based metric models are more effective in assessing the order of relevance of answers similarly to the established standard, when considering the strength and direction of the linear relationship between the assigned scores and the reference scores, the distinction between models that use or do not use a reference answers becomes less distinct.

In relevance tests, the GPT-4 model obtained the highest correlation coefficients across the entire set, demonstrating evident superiority in assigning relevance scores similarly to those of human evaluators. Again, this model is known for its sophisticated ability to understand complex contexts. Therefore, the ability to differentiate what is relevant from what is irrelevant in the answer allowed the model to assign scores more similar to those used by humans. This is shown mainly by the high Pearson coefficient, indicating a strong linear correlation between the relevance scores assigned by GPT-4 and human evaluators.

The GPT-3.5 model also showed high performance compared to other metric models. Its performance was below GPT-4, but with the same highlight on the Pearson coefficient, indicating again a strong linear correlation between the relevance scores it assigned and those by human evaluators. This suggests that, although there is a difference in number of parameters and processing capacity between GPT-3.5 and GPT-4, both models can some how simulate better the human ability to judge the relevance of answers, highlighting in comparison to other models evaluated in the study.

Another model that became notable was the BERTScore, achieving the best result among metric models that use a reference. This metric is originally used to calculate the semantic similarity between texts. In the case of the experiment, it used the precision value returned by the BERTScore model, which focuses on the semantic of the words used in the evaluated answer in relation to the words in the reference answer. This means that for an answer to be considered relevant, the words chosen in the prediction need to have high semantic correspondents in the reference texts. This could be a possible justification for its highlighted performance. In addition to this, this model has the ability to capture the meaning of words, which other models based on lexical overlap do not have, such as ROUGE and BLEU.

The Unities[reference] proposed model also showed considerable performance in the experiments. This model is based on the precision metric, which seeks to determine the relevance score based on the relationship of relevant information in the answer to the total information in it. By using sentences as units of information and employing semantic embeddings derived from the BERT model, the Unities model demonstrates the ability to capture the meaning of the sentences in the answers considering semantic aspects. In terms of relevance, only the Unities[reference] model, which uses a reference answer, obtained highlighted results. The Unities + Google and Unities + Wikipedia models achieved lower results, indicating that documents relevant to the question are considerably less efficient than using a reference answer.

The BLEURT[reference] model also became notable in the experiments with the relevance of answers, showing some effectiveness in assigning relevance scores that correlate well with human evaluations. BLEURT is a text evaluation metric that combines deep learning with a contextualized understanding of language. A possible reason for its performance is related to its construction. BLEURT is trained on a wide dataset of text evaluations, including pairs of texts with human scores, which allows it to capture details in relevance evaluation that simpler models may not detect. This means it is specifically optimized to understand what makes an answer considered relevant by humans.

Another model that became notable was the ROUGE[reference], which measures the overlap of units like n-grams between the evaluated answer and the reference answer. Its efficacy in the relevance experiments can be attributed to its efficient way of capturing lexical overlap, providing a direct measure of how much relevant content is shared between the two texts.

Regarding special metrics, there is a preference among human evaluators to assign higher relevance to the answers generated by GPT-4, as the AlwaysGPT model showed high accuracy, indicating that in 85% of cases, the answer with the highest relevance, according to the human evaluator, is the one produced by the GPT-4 model. Compared to the completeness criterion, human evaluators preferred the answer from the GPT-4 model 97% of the time, which indicates that, although GPT-4 is highly favored by human evaluators in terms of completeness, its preference in terms of relevance is not as dominant. This suggests that while answers generated by GPT-4 are almost universally seen as complete, there are more variations in the perception of their relevance.

The special Length metric revealed an interesting trend in human evaluation: longer answers tend to receive higher relevance scores, with an approximate accuracy of 76% for this metric. At first sight, this may seem counterintuitive, as a more extensive answer could theoretically contain more irrelevant information. However, a possible explanation for this lies in the high quality of the answers evaluated in this experiment. In this context, when long answers contain a lot of relevant information, any irrelevant details become less noticeable and have little impact on the overall relevance of the answer. This is because the abundance of relevant content overshadows the irrelevant parts. Conversely, in shorter answers, any irrelevant information is more noticeable and can negatively impact the answer relevance score.

Tests with different sets of answers showed that relevance metric models had considerable difficulty in evaluating the set of answers produced by the GPT-4 model. In all correlation coefficients, the score was significantly lower. This might indicate that the answers generated by GPT-4 could be characterized by a complexity that challenges the ability of the metric models tested to evaluate relevance. The difficulty of the metric models suggests a possible misalignment between what metric models consider relevant and the judgment of relevance by human evaluators. While humans may assess relevance based on a complex contextual understanding and subjective criteria, metrics might be based on more objective and less flexible criteria, resulting in evaluations that do not fully capture the essence of relevance as perceived by humans. Additionally, this highlights how content generated by IA is becoming more sophisticated, in a point where conventional evaluation tools may not be sufficient.

Finally, compared to the results in the completeness criterion, metric models faced greater difficulty in evaluating the relevance criterion, as, in general, the scores were lower than those for completeness. Additionally, different models became notable in each criterion. The Regression Model was a highlight for completeness but showed inferior results in tests with the relevance criterion, similar to the RankGen and TopicDiversity models. The CosineDistance models also showed low performance compared to others. On the other hand, models like BERTScore[reference], BERTRT[reference], and ROUGE[reference] did not stand out in the completeness criterion but showed competitive results in the relevance criterion.

The variation in the models that becomes notable in each criterion reinforces one of the hypotheses of this research, that there is no single approach to evaluate all aspects of the answers, with specific approaches being necessary for each criterion to be evaluated. Each metric model has its own strengths and limitations depending on the evaluation criteria.

The lower performance in relevance, compared to completeness, may be related to the subjectivity of the criterion, which is shown in a lower annotation agreement score among annotators, as presented in Section 5.4.1. This reflects the challenge in evaluating the relevance of an answer, which is influenced by personal interpretations, the context of the question, and details of each evaluator's understanding of the content. The completeness of an answer should be a more objective criterion where it is verified whether all the relevant information of a question has been covered, which allows a more direct and possibly concordant evaluation between the different humans evaluators. On the other hand, relevance may require a deeper analysis of the relevance of the answer content in relation to the specific question, along with the issue of subjectivity, which can vary significantly between individuals.

7.3 Final Considerations on the Results

The results of the experiments with the completeness and relevance criteria revealed important observations about the performance of various metric models. By breaking down the results by criterion and type of answer, it was possible to discern important details that help to better understand how each model operates and what their specific strengths and limitations are.

The experiments highlighted the competence of models such as GPT-4 in replicating human judgments, especially in terms of completeness. This ability of GPT-4 to align its evaluations with human perceptions suggests a significant advancement in the automated understanding of complex texts, although there are still challenges in terms of relevance. The results also emphasized the importance of specialized metrics, tailored to specific evaluation criteria, such as BERTScore[reference] for relevance and the Regression Model for completeness, reinforcing the idea that different aspects of answer evaluation demand distinct approaches.

The experiments revealed the remarkable ability of the proposed models, particularly for the completeness criterion, like the Regression Model, which, in various aspects, overcome GPT-4 in terms of performance. Its effectiveness, along with the fact that it is economically more viable and independent of reference answers, highlights the value of developing metrics adjusted to specific evaluation criteria. Moreover, the Unities + Google model proved to be a promising model, providing competitive performance without the need for direct reference data.

The detailed analysis of special metrics, such as AlwaysGPT and Length, offered insights into the trends and preferences of human evaluators, like the propensity in favor of answers generated by GPT-4 and the association between the length of the answer and its perceived completeness or relevance. These observations underline the complexity of the evaluation task and the need to consider human and contextual factors in interpreting the results.

When examining the sets of answers generated by humans and GPT-4 separately, the variability in the performance of metrics became evident, highlighting the impact of the type of answer on the difficulty of evaluation. Therefore, with regard to assessing relevance, the models used to measure performance faced significant challenges in evaluating answers produced by GPT-4. This suggests that there is a gap between the complexity of answers generated by AI and the current solutions' capability to evaluate relevance.

This chapter provided a comprehensive overview of the challenges and considerations in evaluating answers based on completeness and relevance, illustrating the importance of adaptive approaches in the search for metrics that aims to replicate human judgment. The results highlight both the potential and limitations of current solutions, pointing the way for future research in optimizing metric models and exploring new methodologies for automatic text evaluation. In response to the rapid development of AI technologies, the future demands innovative assessment strategies that can adapt to and measure the specialized features of such texts.

8 CONCLUSION

This thesis proposes the development and validation of metrics designed to assess criteria for completeness and relevance of long answers generated by QA systems. Through these metrics, it is possible to verify how complete and relevant the answers generated by the system are, thereby allowing adjustments according to the specificities of each system. Therefore, this work begins with the contribution of a systematic review of non-factoid QA systems, highlighting the main methods, tasks, datasets, evaluation strategies, and results obtained by the analyzed systems.

Another important contribution was the creation of a dataset of answers evaluated by humans for completeness and relevance criteria, enabling new researchers to develop and test their evaluation approaches of such criteria and compare their results. The work also proposed three distinct approaches to measure the completeness and relevance of long answers, which do not require a reference answer, through: a prompt strategy with GPT models; a model that segments information into discrete units and uses formulas analogous to precision and recall to determine relevance and completeness, respectively; and regression models trained with synthetic data to assign completeness and relevance scores.

Through experiments comparing different metrics used for evaluating long answers, along with the proposed metric models, this work presented a study that shows how much these metrics correlate with human evaluation for the criteria of completeness and relevance, highlighting a high correlation for the prompt approach with the GPT-4 model, a high performance of the proposed metric using a regression model for the completeness criterion, and demonstrating that metrics that do not require reference answers are competitive, as well as showing better performance in different scenarios.

The synthesis of these contributions brings advancements in the evaluation of QA systems that provide long answers, which can be improved by specifically focusing on the criteria of completeness and relevance. The metrics developed in this work propose a refined and detailed approach, essential for understanding the effectiveness of these systems beyond conventional metrics like BLEU and ROUGE, which often do not capture the semantic details and linguistic flexibility of the answers. The creation of an annotated dataset for these specific criteria and the development of metric models that do not depend on reference answers represent advancements to overcome the limitations of traditional evaluation methods, such as the dependency on “golden” answers and the high demand for human effort. This study seeks to provide a foundation for future research, aimed at the continuous improvement of QA systems in providing informative answers to users.

8.1 Limitations and Future Works

This research contains some limitations that need to be considered and that serve as starting points for future research. Being an initial effort with limited resources, the dataset built and used in the experiments focuses specifically on “instruction” type questions. Future work could cover other types of questions, such as “descriptive”, “comparativ”, and “explanatory”, which also require long answers and can be evaluated with the proposed criteria. Moreover, the dataset specializes in computer science, which is useful to ensure the expertise of the evaluators but limits the generalization of the results. Subsequent studies should include other areas of knowledge, increasing thematic diversity and providing a more comprehensive understanding in different domains.

The constructed dataset consists of 106 questions and 212 answers. The relatively small number of questions and answers suggests a need for expansion to enable more robust tests, such as eventually, data for training and validation of models.

Regarding the experiments, the aim was to utilize the main metrics that allow the evaluation of long answers, especially those that have some relation to the criteria of completeness and relevance. However, there are many other metrics that could be evaluated and may potentially have a high correlation with human annotations. Also, the dependence on reference answers for many of the used metrics can be mitigated. Investigating the use of different types of reference information, such as academic documents or specialized databases, could offer new approaches and enhancements of existing metrics.

A certain difficulty was observed on the part of human annotators in assigning scores for the relevance criterion, highlighting the need to develop more effective annotation strategies. Future investigations could explore new annotation methods or adaptations of existing criteria to facilitate the assignment of relevance.

Additionally, it is important to consider a characteristic related to the quality of the answers within the dataset used for testing. The answers included are primarily those most upvoted on Reddit or generated by GPT-4, which implies a generally high quality. Consequently, there is a deficiency in the dataset concerning the representation of answers with low relevance, which could impact the evaluation of relevance. Future iterations of the dataset should consider incorporating a broader range of answer qualities, including those with lower relevance, to provide a more comprehensive test environment for the tested models.

Regarding generative LLMs, this study only used the GPT-3.5 and GPT-4 models with a single prompt strategy. There are various other models that could be used and new prompt strategies could bring performance improvements. In relation to the proposed method that segments the answer into information units, this study used only one method of dividing the answer into sentences. However, other methods could be tested, such as Open Information Extraction to divide the text into information tuples, which could enhance the quality of representing answers in information units.

REFERENCES

ABDEL-NABI, H.; AWAJAN, A.; ALI, M. Z. Deep learning-based question answering: a survey. **Knowledge and Information Systems**, v. 65, n. 4, p. 1399–1485, 04 2023. ISSN 0219-3116. Available from Internet: <https://doi.org/10.1007/s10115-022-01783-5>.

ACHIAM, J. et al. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

AGICHTEIN, E. et al. Overview of the trec 2015 liveqa track. In: **TREC**. [S.l.: s.n.], 2015.

AHARONI, R. et al. Multilingual summarization with factual consistency evaluation. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Findings of the Association for Computational Linguistics: ACL 2023**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 3562–3591. Available from Internet: <https://aclanthology.org/2023.findings-acl.220>.

AL-OMARI, H.; DUWAIRI, R. So2al-wa-gwab: A new arabic question-answering dataset trained on answer extraction models. **ACM Trans. Asian Low-Resour. Lang. Inf. Process.**, Association for Computing Machinery, New York, NY, USA, v. 22, n. 8, aug 2023. ISSN 2375-4699. Available from Internet: <https://doi.org/10.1145/3605550>.

AMPLAYO, R. K. et al. **SMART: Sentences as Basic Units for Text Evaluation**. 2022.

BAE, K.; KO, Y. Efficient question classification and retrieval using category information and word embedding on cQA services. **Journal of Intelligent Information Systems**, v. 53, n. 1, p. 27–49, 2019. ISSN 15737675. Available from Internet: <https://doi.org/10.1007/s10844-019-00556-x>.

BAJAJ, P. et al. **MS MARCO: A Human Generated MAchine Reading COmprehension Dataset**. 2018.

BANERJEE, S.; LAVIE, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: GOLDSTEIN, J. et al. (Ed.). **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Available from Internet: <https://aclanthology.org/W05-0909>.

BARRY, C. L.; SCHAMBER, L. Users' criteria for relevance evaluation: A cross-situational comparison. **Information Processing Management**, v. 34, n. 2, p. 219–236, 1998. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457397000782>.

Ben Abacha, A.; ZWEIGENBAUM, P. Means: A medical question-answering system combining nlp techniques and semantic web technologies. **Information Processing & Management**, v. 51, n. 5, p. 570–594, 2015. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457315000515>.

BIDGOLY, A. J.; AMIRKHANI, H.; BARADARAN, R. Clustering-based sequence to sequence model for generative question answering in a low-resource language. **ACM Trans. Asian Low-Resour. Lang. Inf. Process.**, Association for Computing Machinery, New York, NY, USA, v. 22, n. 2, dec 2022. ISSN 2375-4699. Available from Internet: <https://doi.org/10.1145/3563036>.

BLOOMA, M. J.; CHUA, A. Y. K.; GOH, D. H.-L. A predictive framework for retrieving the best answer. In: **Proceedings of the 2008 ACM Symposium on Applied Computing**. New York, NY, USA: Association for Computing Machinery, 2008. (SAC '08), p. 1107–1111. ISBN 9781595937537. Available from Internet: <https://doi.org/10.1145/1363686.1363944>.

BOLOTOVA-BARANOVA, V. et al. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 5291–5314. Available from Internet: <https://aclanthology.org/2023.acl-long.290>.

BOLOTOVA, V. et al. **A Non-Factoid Question-Answering Taxonomy**. New York, NY, USA: Association for Computing Machinery, 2022. 1196–1207 p. (SIGIR '22). Available from Internet: <https://doi.org/10.1145/3477495.3531926>.

BONDARENKO, A. et al. Comparative web search questions. In: **Proceedings of the 13th International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2020. (WSDM '20), p. 52–60. ISBN 9781450368223. Available from Internet: <https://doi.org/10.1145/3336191.3371848>.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. 2020.

CAMBAZOGLU, B. B. et al. Quantifying human-perceived answer utility in non-factoid question answering. In: **Proceedings of the 2021 Conference on Human Information Interaction and Retrieval**. New York, NY, USA: Association for Computing Machinery, 2021. (CHIIR '21), p. 75–84. ISBN 9781450380553. Available from Internet: <https://doi.org/10.1145/3406522.3446028>.

CAO, Y. et al. Askhermes: An online question answering system for complex clinical questions. **Journal of Biomedical Informatics**, v. 44, n. 2, p. 277–288, 2011. ISSN 1532-0464. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1532046411000062>.

CEGIN, J.; SIMKO, J.; BRUSILOVSKY, P. ChatGPT to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 1889–1905. Available from Internet: <https://aclanthology.org/2023.emnlp-main.117>.

CELIKYILMAZ, A.; CLARK, E.; GAO, J. **Evaluation of Text Generation: A Survey**. 2021.

CHALI, Y.; HASAN, S. A.; MOJAHID, M. A reinforcement learning formulation to the complex question answering problem. **Information Processing & Management**, v. 51, n. 3, p. 252–272, 2015. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457315000035>.

CHAYBOUTI, S.; SAGHE, A.; SHABOU, A. **EfficientQA : a RoBERTa Based Phrase-Indexed Question-Answering System**. 2021.

CHEN, Y.; EGER, S. MENLI: Robust Evaluation Metrics from Natural Language Inference. **Transactions of the Association for Computational Linguistics**, v. 11, p. 804–825, 07 2023. ISSN 2307-387X. Available from Internet: <https://doi.org/10.1162/tacl_a_00576>.

CHU, S. K. W. et al. Quality and clarity of health information on qa sites. **Library Information Science Research**, v. 40, n. 3, p. 237–244, 2018. ISSN 0740-8188. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0740818817303390>.

CLARK, K. et al. **ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators**. 2020.

COHEN, D.; YANG, L.; CROFT, W. B. WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. **41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018**, p. 1165–1168, 2018.

CORBIN, J.; STRAUSS, A. **Basics of qualitative research: Techniques and procedures for developing grounded theory**. [S.l.]: Sage publications, 2014.

CORTES, E. et al. An empirical comparison of question classification methods for question answering systems. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 5408–5416. ISBN 979-10-95546-34-4. Available from Internet: <https://www.aclweb.org/anthology/2020.lrec-1.665>.

CORTES, E. G. et al. A systematic review of question answering systems for non-factoid questions. **Journal of Intelligent Information Systems**, v. 58, n. 3, p. 453–480, 2022. ISSN 1573-7675. Available from Internet: <https://doi.org/10.1007/s10844-021-00655-8>.

CORTES, E. G.; WOLOSZYN, V.; BARONE, D. A. C. When, where, who, what or why? a hybrid model to question answering systems. In: VILLAVICENCIO, A. et al. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2018. p. 136–146. ISBN 978-3-319-99722-3.

COSTA, J. O.; KULKARNI, A. Leveraging knowledge graph for open-domain question answering. In: **2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)**. [S.l.: s.n.], 2018. p. 389–394.

DARVISHI, K. et al. Pquad: A persian question answering dataset. **Computer Speech Language**, v. 80, p. 101486, 2023. ISSN 0885-2308. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0885230823000050>.

DELONE, W. H.; MCLEAN, E. R. The delone and mclean model of information systems success: A ten-year update. **Journal of Management Information Systems**, Routledge, v. 19, n. 4, p. 9–30, 2003. Available from Internet: <https://doi.org/10.1080/07421222.2003.11045748>.

DENG, Y. et al. Nonfactoid question answering as query-focused summarization with graph-enhanced multihop inference. **IEEE Transactions on Neural Networks and Learning Systems**, p. 1–15, 2023.

DENYER, D.; TRANFIELD, D. Producing a systematic review. **Sage Publications Ltd**, Sage Publications Ltd, p. 671—689, 2009.

DEUTSCH, D.; BEDRAX-WEISS, T.; ROTH, D. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. **Transactions of the Association for Computational Linguistics**, v. 9, p. 774–789, 08 2021. ISSN 2307-387X. Available from Internet: <https://doi.org/10.1162/tacl_a_00397>.

DEUTSCH, D.; ROTH, D. Understanding the extent to which content quality metrics measure the information quality of summaries. In: BISAZZA, A.; ABEND, O. (Ed.). **Proceedings of the 25th Conference on Computational Natural Language Learning**. Online: Association for Computational Linguistics, 2021. p. 300–309. Available from Internet: <https://aclanthology.org/2021.conll-1.24>.

DEUTSCH, D.; ROTH, D. Benchmarking answer verification methods for question answering-based summarization evaluation metrics. In: MURESAN, S.; NAKOV, P.; VILLAVICENCIO, A. (Ed.). **Findings of the Association for Computational Linguistics: ACL 2022**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 3759–3765. Available from Internet: <https://aclanthology.org/2022.findings-acl.296>.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. **CoRR**, abs/1810.04805, 2018. Available from Internet: <http://arxiv.org/abs/1810.04805>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DIMITRAKIS, E.; SGONTZOS, K.; TZITZIKAS, Y. A survey on question answering systems over linked data and documents. **Journal of Intelligent Information Systems**, v. 55, p. 233 – 259, 2019.

DYBå, T.; DINGSøYR, T. Empirical studies of agile software development: A systematic review. **Information and Software Technology**, v. 50, n. 9, p. 833–859, 2008. ISSN 0950-5849. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0950584908000256>.

FABBRI, A. et al. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In: CARPUAT, M.; MARNEFFE, M.-C. de; RUIZ, I. V. M. (Ed.). **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Seattle, United States: Association for Computational Linguistics, 2022. p. 2587–2601. Available from Internet: <https://aclanthology.org/2022.naacl-main.187>.

FAGGIOLI, G. et al. Perspectives on large language models for relevance judgment. In: **Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2023. (ICTIR '23), p. 39–50. ISBN 9798400700736. Available from Internet: <https://doi.org/10.1145/3578337.3605136>.

FAN, A. et al. Eli5: Long form question answering. In: **Proceedings of ACL 2019**. [S.l.: s.n.], 2019.

FICHMAN, P. A comparative assessment of answer quality on four question answering sites. **Journal of Information Science**, Sage Publications Sage UK: London, England, v. 37, n. 5, p. 476–486, 2011.

FRICKÉ, M. Information using likeness measures. **Journal of the American Society for Information Science**, Wiley Online Library, v. 48, n. 10, p. 882–892, 1997.

GAO, M. et al. **LLM-based NLG Evaluation: Current Status and Challenges**. 2024.

GEHRMANN, S.; CLARK, E.; SELLAM, T. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. **Journal of Artificial Intelligence Research**, v. 77, p. 103–166, 2023.

GERSTENBERGER, C. et al. Instant annotations – applying NLP methods to the annotation of spoken language documentation corpora. In: TYERS, F. M. et al. (Ed.). **Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages**. St. Petersburg, Russia: Association for Computational Linguistics, 2017. p. 25–36. Available from Internet: <https://aclanthology.org/W17-0604>.

HE, P.; GAO, J.; CHEN, W. **DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing**. 2021.

HERMJAKOB, U.; ECHIHABI, A.; MARCU, D. Natural language based reformulation resource and web exploitation for question answering. In: CITESEER. **Proceedings of TREC**. [S.l.], 2002. v. 11.

HIGGINS, J. et al. (Ed.). **Cochrane Handbook for Systematic Reviews of Interventions**. 2nd. ed. Australia: Wiley-Blackwell, 2019. ISBN 9781119536628.

HONOVICH, O. et al. TRUE: Re-evaluating factual consistency evaluation. In: CARPUAT, M.; MARNEFFE, M.-C. de; RUIZ, I. V. M. (Ed.). **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Seattle, United States: Association for Computational Linguistics, 2022. p. 3905–3920. Available from Internet: <https://aclanthology.org/2022.naacl-main.287>.

HU, N. et al. An empirical study of pre-trained language models in simple knowledge graph question answering. **World Wide Web**, Springer, v. 26, n. 5, p. 2855–2886, 2023.

ISABELLE, P.; CHERRY, C.; FOSTER, G. **A Challenge Set Approach to Evaluating Machine Translation**. 2017.

JIANG, Z.; CHI, C.; ZHAN, Y. Research on medical question answering system based on knowledge graph. **IEEE Access**, v. 9, p. 21094–21101, 2021.

JR, B. F. G. et al. Baseball: an automatic question-answerer. In: ACM. **Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference**. [S.l.], 1961. p. 219–224.

KE, P. et al. DecompEval: Evaluating generated texts as unsupervised decomposed question answering. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 9676–9691. Available from Internet: <https://aclanthology.org/2023.acl-long.539>.

KEELER, M. Crowdsourced knowledge: Peril and promise for conceptual structures research. In: ANDREWS, S. et al. (Ed.). **Conceptual Structures for Discovering Knowledge**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 131–144. ISBN 978-3-642-22688-5.

KHAN, K. S. et al. Five steps to conducting a systematic review. **Journal of the Royal Society of Medicine**, v. 96, n. 3, p. 118–121, 2003. PMID: 12612111. Available from Internet: <https://doi.org/10.1177/014107680309600304>.

KHILJI, A. F. U. R. et al. Cookingqa: answering questions and recommending recipes based on ingredients. **Arabian Journal for Science and Engineering**, Springer, v. 46, n. 4, p. 3701–3712, 2021.

KHUSHHAL, S. et al. Question retrieval using combined queries in community question answering. **Journal of Intelligent Information Systems**, v. 55, p. 307 – 327, 2020. Available from Internet: <https://doi.org/10.1007/s10844-020-00612-x>.

KIM, G. et al. Donut: Document understanding transformer without OCR. **CoRR**, abs/2111.15664, 2021. Available from Internet: <https://arxiv.org/abs/2111.15664>.

KIM, S.; OH, S. Users' relevance criteria for evaluating answers in a social q&a site. **Journal of the American society for information science and technology**, Wiley Online Library, v. 60, n. 4, p. 716–727, 2009.

KIM, S.-E. et al. Effects of tourism information quality in social media on destination image formation: The case of sina weibo. **Information Management**, v. 54, n. 6, p. 687–702, 2017. ISSN 0378-7206. Smart Tourism: Traveler, Business, and Organizational Perspectives. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0378720617301295>.

KIM, W.; SON, B.; KIM, I. **ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision**. 2021.

KOCMI, T.; FEDERMANN, C. **GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4**. 2023.

KOCMI, T.; FEDERMANN, C. Large language models are state-of-the-art evaluators of translation quality. In: NURMINEN, M. et al. (Ed.). **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**. Tampere, Finland: European Association for Machine Translation, 2023. p. 193–203. Available from Internet: <https://aclanthology.org/2023.eamt-1.19>.

KODRA, L.; KAJO, E. Question answering systems: A review on present developments, challenges and trends. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 9, p. 217–224, 2017. ISSN 2158107X.

KOLOMIYETS, O.; MOENS, M. F. A survey on question answering technology from an information retrieval perspective. **Information Sciences**, Elsevier Inc., v. 181, n. 24, p. 5412–5434, 12 2011. ISSN 00200255.

KRISHNA, K. et al. Rankgen: Improving text generation with large ranking models. In: **Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2022.

LAN, Z. et al. **ALBERT: A Lite BERT for Self-supervised Learning of Language Representations**. 2020.

LEWIS, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. **arXiv preprint arXiv:1910.13461**, 2019.

LI, B. et al. Hierarchical sliding inference generator for question-driven abstractive answer summarization. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 41, n. 1, jan 2023. ISSN 1046-8188. Available from Internet: <https://doi.org/10.1145/3511891>.

LI, L.; ZHANG, C.; HE, D. Factors influencing the importance of criteria for judging answer quality on academic social q&a platforms. **Aslib Journal of Information Management**, Emerald Publishing Limited, v. 72, n. 6, p. 887–907, 2020.

LI, L. et al. Researchers' judgment criteria of high-quality answers on academic social q&a platforms. **Online Information Review**, Emerald Publishing Limited, v. 44, n. 3, p. 603–623, 2020.

LI, L. et al. Investigating factors for assessing the quality of academic user-generated content on social media. In: **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020**. New York, NY, USA: Association for Computing Machinery, 2020. (JCDL '20), p. 511–512. ISBN 9781450375856. Available from Internet: <https://doi.org/10.1145/3383583.3398588>.

LI, R.; PATEL, T.; DU, X. **PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations**. 2023.

LI, W.; LI, W.; WU, Y. A unified model for document-based question answering based on human-like reading strategy. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 32, n. 1, Apr 2018. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/11316>.

LI, X.; ROTH, D. Learning question classifiers. In: **COLING 2002: The 19th International Conference on Computational Linguistics**. [S.l.: s.n.], 2002.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Available from Internet: <https://aclanthology.org/W04-1013>.

LIN, C. Y.; WU, Y.-H.; CHEN, A. L. P. Selecting the most helpful answers in online health question answering communities. **Journal of Intelligent Information Systems**, v. 57, n. 2, p. 271–293, 10 2021. Available from Internet: <https://doi.org/10.1007/s10844-021-00640-1>.

LIU, Y. et al. **G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment**. 2023.

LIU, Y. et al. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019.

LIU, Y. et al. A Survey on Frameworks and Methods of Question Answering. **Proceedings - 2016 3rd International Conference on Information Science and Control Engineering, ICISCE 2016**, IEEE, p. 115–119, 2016.

LYU, C. et al. Improving unsupervised question answering via summarization-informed question generation. In: MOENS, M.-F. et al. (Ed.). **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 4134–4148. Available from Internet: <https://aclanthology.org/2021.emnlp-main.340>.

MADABUSHI, H. T.; LEE, M. High accuracy rule-based question classification using question syntax and semantics. In: MATSUMOTO, Y.; PRASAD, R. (Ed.). **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 1220–1230. Available from Internet: <https://aclanthology.org/C16-1116>.

MALAVIYA, C. et al. **ExpertQA: Expert-Curated Questions and Attributed Answers**. 2024.

MAO, R. et al. **GPTEval: A Survey on Assessments of ChatGPT and GPT-4**. 2023.

MIN, S. et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 12076–12100. Available from Internet: <https://aclanthology.org/2023.emnlp-main.741>.

NAKANO, R. et al. **WebGPT: Browser-assisted question-answering with human feedback**. 2022.

NIE, Y.-p. et al. Attention-based encoder-decoder model for answer selection in question answering. **Frontiers of Information Technology & Electronic Engineering**, v. 18, n. 4, p. 535–544, 2017. ISSN 2095-9230. Available from Internet: <https://doi.org/10.1631/FITEE.1601232>.

NORASET, T.; LOWPHANSIRIKUL, L.; TUAROB, S. Wabiqa: A wikipedia-based thai question-answering system. **Information Processing & Management**, v. 58, n. 1, p. 102431, 2021. ISSN 0306-4573.

OMAR, R. et al. **ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots**. 2023.

OPENAI et al. **GPT-4 Technical Report**. 2024.

OSTYAKOVA, L. et al. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In: STOYANCHEV, S. et al. (Ed.). **Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and**

Dialogue. Prague, Czechia: Association for Computational Linguistics, 2023. p. 242–254. Available from Internet: <https://aclanthology.org/2023.sigdial-1.23>.

OUZZANI, M. et al. Rayyan—a web and mobile app for systematic reviews. **Systematic reviews**, Springer, v. 5, n. 1, p. 210, 2016.

PAPADAKIS, M.; TZITZIKAS, Y. Answering keyword queries through cached subqueries in best match retrieval models. **J. Intell. Inf. Syst.**, Kluwer Academic Publishers, USA, v. 44, n. 1, p. 67–106, feb. 2015. ISSN 0925-9902. Available from Internet: <https://doi.org/10.1007/s10844-014-0330-7>.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Available from Internet: <https://doi.org/10.3115/1073083.1073135>.

PEREIRA, A. et al. Systematic review of question answering over knowledge bases. **IET Software**, Wiley Online Library, v. 16, n. 1, p. 1–13, 2022.

PITHYAACHARIYAKUL, C.; KULKARNI, A. Automated question answering system for community-based questions. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 32, n. 1, Apr. 2018. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/12159>.

QIU, Y. et al. Hierarchical query graph generation for complex question answering over knowledge graph. In: **Proceedings of the 29th ACM International Conference on Information & Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2020. (CIKM '20), p. 1285–1294. ISBN 9781450368599. Available from Internet: <https://doi.org/10.1145/3340531.3411888>.

RAJPURKAR, P.; JIA, R.; LIANG, P. **Know What You Don't Know: Unanswerable Questions for SQuAD**. 2018.

SACHAN, D. S. et al. **End-to-End Training of Neural Retrievers for Open-Domain Question Answering**. 2021.

SANH, V. et al. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. 2020.

SASIKUMAR, U.; SINDHU, L. A survey of natural language question answering system. **International Journal of Computer Applications**, Foundation of Computer Science, v. 108, n. 15, p. 975–8887, 2014.

SEERS, K. Qualitative data analysis. **Evidence-based nursing**, Royal College of Nursing, v. 15, n. 1, p. 2–2, 2012.

SOARES, M. A. C.; PARREIRAS, F. S. A literature review on question answering techniques, paradigms and systems. **Journal of King Saud University - Computer and Information Sciences**, King Saud bin Abdulaziz University, v. 32, n. 6, p. 635–646, 7 2020. ISSN 22131248.

SPECIA, L.; SCARTON, C.; PAETZOLD, G. H. Quality estimation for machine translation. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 11, n. 1, p. 1–162, 2018.

STVILIA, B. et al. Information quality work organization in wikipedia. **Journal of the American society for information science and technology**, Wiley Online Library, v. 59, n. 6, p. 983–1001, 2008.

SULEM, E.; ABEND, O.; RAPPOPORT, A. BLEU is not suitable for the evaluation of text simplification. In: RILOFF, E. et al. (Ed.). **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 738–744. Available from Internet: <https://aclanthology.org/D18-1081>.

SURDEANU, M.; CIARAMITA, M.; ZARAGOZA, H. Learning to rank answers on large online QA collections. In: **ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference**. [S.l.: s.n.], 2008. p. 719–727. ISBN 9781932432046.

SURYANTO, M. A. et al. Quality-aware collaborative question answering: methods and evaluation. In: **Proceedings of the Second ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2009. (WSDM '09), p. 142–151. ISBN 9781605583907. Available from Internet: <https://doi.org/10.1145/1498759.1498820>.

TRANFIELD, D.; DENYER, D.; SMART, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. **British journal of management**, Wiley Online Library, v. 14, n. 3, p. 207–222, 2003.

TöRNBERG, P. **ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning**. 2023.

WANG, X. et al. Japanese how-to tip machine reading comprehension by multi-task learning based on generative model. In: EKŠTEIN, K.; PÁRTL, F.; KONOPÍK, M. (Ed.). **Text, Speech, and Dialogue**. Cham: Springer Nature Switzerland, 2023. p. 3–14. ISBN 978-3-031-40498-6.

WU, Y. et al. Leveraging social Q&A collections for improving complex question answering. **Computer Speech and Language**, Elsevier Ltd, v. 29, n. 1, p. 1–19, 2015. ISSN 10958363. Available from Internet: <http://dx.doi.org/10.1016/j.csl.2014.06.001>.

XIA, J.; WU, C.; YAN, M. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. In: **Proceedings of the 28th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2019. (CIKM '19), p. 2393–2396. ISBN 9781450369763. Available from Internet: <https://doi.org/10.1145/3357384.3358165>.

XU, F. et al. A critical evaluation of evaluations for long-form question answering. In: **Association of Computational Linguistics**. [S.l.: s.n.], 2023.

YAN, Z.; ZHOU, J. Optimal answerer ranking for new questions in community question answering. **Information Processing & Management**, v. 51, n. 1, p. 163–178, 2015. ISSN 0306-4573.

YANG, M. et al. Investigating the transferring capability of capsule networks for text classification. **Neural Networks**, v. 118, p. 247 – 261, 2019. ISSN 0893-6080. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S089360801930187X>.

YASUNAGA, M. et al. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. **arXiv preprint arXiv:2104.06378**, 2021.

YOGISH, D.; MANJUNATH, T. N.; HEGADI, R. S. Survey on trends and methods of an intelligent answering system. **International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017**, v. 2018-Janua, p. 346–353, 2018.

YUAN, W.; NEUBIG, G.; LIU, P. **BARTScore: Evaluating Generated Text as Text Generation**. 2021.

ZHANG*, T. et al. **BERTScore: Evaluating Text Generation with BERT**. 2020. Available from Internet: <https://openreview.net/forum?id=SkeHuCVFDr>.

ZHANG, Z.; YANG, J.; ZHAO, H. **Retrospective Reader for Machine Reading Comprehension**. 2020.

ZHU, F. et al. **Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering**. 2021.

ZIRUI, C. et al. Survey of open-domain knowledge graph question answering. **Journal of Frontiers of Computer Science & Technology**, v. 15, n. 10, 2021.

Appendix A: Systematic Review - Papers Features

ID	Year	Tasks	Datasets	Language	Question Type	Domain of Knowledge	Evaluation Metrics	Knowledge Source Type
P1	2020	Candidate Answers Ranking	ANTIQUE	English	-	Open-Domain	P@k, MAP, MRR, NDCG	Documents
P2	2020	Candidate Answers Extraction	SimpleQuestions	English	Definition, Confirmation, Factoid	Open-Domain	Accuracy	Knowledge Graph
P3	2020	Question Classification	Created by Authorts	Russian	Comparison, Opinion, Factoid	Open-Domain	F1	-
P4	2020	Question Classification, Document Retrieval, Passage Extraction, Candidate Answers Extraction	BioASQ	English	Confirmation, Factoid	Health	F1, Accuracy, MRR	Documents
P5	2020	Candidate Answer Extraction	Created by Authorts	Japonese	How	Open-Domain	Accuracy	Documents
P6	2019	Candidate Answers Ranking	L6 - Yahoo! Answers Comprehensive QA	English	-	Open-Domain	P@1, MRR	Documents
P7	2019	Document Retrieval	HealthQA	English	Why, How, Definition, Factoid, Confirmation	Health	MRR, Recall@K, DRMM, KNRM, aNMM, Duet, MatchPyramid	Documents
P8	2019	Question Reformulation	Created by Authorts	Chinese	-	Geographical	F1, MRR	Documents, Knowledge Graph
P9	2019	Candidate Answers Ranking	SemEval-2015, 2016 and 2017	English	-	Open-Domain	F1, Accuracy	Documents, Answer List
P10	2019	Candidate Answers Ranking	L5 - Yahoo! Answers Manner Questions	English	How	Open-Domain	P@k, MRR	Answer List
P11	2019	Candidate Answers Ranking	Yahoo! Answers	English	Comparison	Open-Domain	P@k, MAP	Web, Answer List
P12	2019	Candidate Answers Extraction	Created by Authorts	English	Comparison	E-Commerce	P@k	Documents, Web, Knowledge Graph
P13	2018	Candidate Answers Extraction	WikiPassageQA	English	-	Open-Domain	MAP, MRR, P@k, NDCG	Documents
P14	2018	Candidate Answers Extraction	WebAP, MSMARCO	English	Factoid	Open-Domain	Accuracy, Rouge	Documents
P15	2018	Question Reformulation, Candidate Answers Extraction, Candidate Answers Ranking	LiveQA TREC	English	Factoid	Open-Domain	Manual	Web, Knowledge Graph
P16	2018	Candidate Answers Extraction	Yahoo! Answers	English	-	Open-Domain	Rouge	Web, Answer List
P17	2018	Candidate Answers Extraction, Candidate Answers Ranking	FiQA, InsuranceQA	English	Opinion	Open-Domain, Insurance	P@k, MRR, NDCG	Documents, Answer List
P18	2018	Candidate Answers Ranking	InsuranceQA, FiQA	English	Opinion	Open-Domain, Financial, Insurance	MAP, MRR	Documents, Answer List
P19	2018	Question Reformulation, Document Retrieval, Candidate Answers Extraction, Candidate Answers Ranking	LiveQA TREC	English	Factoid	Open-Domain	Manual	Web, Answer List
P20	2018	Candidate Answers Ranking	Created by Authorts	English	Factoid	Open-Domain	MRR	Web
P21	2018	Candidate Answers Ranking	WebAP, nfl6, WikiPassageQA	English	-	Open-Domain	MRR, NDCG	Answer List
P22	2018	Candidate Answers Ranking	LiveQA TREC, InsuranceQA	English	-	Open-Domain	MAP, MRR	Answer List
P23	2018	Candidate Answers Extraction, Candidate Answers Ranking	GOV2, ClueWeb09B	English	-	Open-Domain	P@k, MAP, MRR, NDCG	Documents
P24	2017	Question Reformulation, Candidate Answers Extraction, Candidate Answers Ranking	L6 - Yahoo! Answers Comprehensive QA, LiveQA TREC	English	-	Open-Domain	MAP, MRR, NDCG	Web
p25	2017	Candidate Answers Ranking	Yahoo! Answers, LiveQA TREC	English	-	Open-Domain	P@1, MRR, Manual	Web
P26	2017	Candidate Answers Ranking	Created by Authorts	English, Chinese	-	Open-Domain	Accuracy	-
P27	2017	Candidate Answers Ranking	InsuranceQA, AgricultureQA	English, Chinese	-	Insurance, Agriculture	Accuracy	Documents
P28	2017	Question Reformulation, Document Retrieval, Passage Extraction, Candidate Answers Extraction, Candidate Answers Ranking	L6 - Yahoo! Answers Comprehensive QA, TREC-QA	English	-	Open-Domain	MAP, MRR, NDCG	Web
P29	2017	Candidate Answers Extraction	Yahoo! Answers	English	-	Open-Domain	Rouge	Answer List
P30	2017	Question Reformulation	WebAP	English	-	Open-Domain	P@k, MRR, NDCG	Documents
P31	2016	Candidate Answers Ranking	L6 - Yahoo! Answers Comprehensive QA	English	How	Open-Domain	P@1, MRR	Documents
P32	2016	Candidate Answers Extraction, Candidate Answers Ranking	WebAP	English	-	Open-Domain	MRR, NDCG, P@10	Web
P33	2016	Candidate Answers Ranking	TREC-QA, AQUAINT, AQUAINT-2	English	Definition, Factoid	Open-Domain	Accuracy	Documents, Web
P34	2016	Question Classification	Created by Authorts	Indonesian	Comparison	Open-Domain	Accuracy	Web
P35	2016	Candidate Answers Extraction	Created by Authorts	Arabic	Why	Open-Domain	c@1	Documents
P36	2016	Candidate Answers Ranking	Created by Authorts	Chinese	-	Tourism	P@k, MRR	Documents
P37	2016	Candidate Answers Extraction	WebAP	English	Definition	Open-Domain	P@k, NDCG, Rouge	Documents
P38	2016	Candidate Answers Ranking	SemEval-2016	English	-	Open-Domain	MAP, MRR	Answer List
P39	2016	Candidate Answers Ranking	Yahoo! Answers	English	-	Open-Domain	P@k, MRR	Answer List
P40	2016	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	MCTest	English	-	Open-domain	MAP, MRR, Accuracy	Documents
P41	2016	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	SQuAd	English	-	Open-Domain	F1, Exact match	Documents
P42	2015	Question Classification	NTCIR	Chinese	-	Open-Domain	F1	Documents, Web
P43	2015	Candidate Answers Ranking	Created by Authorts	English	-	Insurance	Accuracy	Documents
P44	2015	Candidate Answers Ranking	Yahoo! Answers	English	-	Open-Domain	MAP, MRR	Answer List
P45	2015	Candidate Answers Ranking	TREC-QA, AQUAINT	English	-	Open-Domain	P@k, MAP, MRR	Documents
P46	2015	Passage Ranking	WebAP	English	-	Open-Domain	P@k, MRR, NDCG	Web

ID	Year	Tasks	Datasets	Language	Question Type	Domain of Knowledge	Evaluation Metrics	Knowledge Source Type
P47	2014	Candidate Answers Ranking	Yahoo! Answers, Biology Textbook Corpus (Bio)	English	Why, How	Restrict-Domain, Biology	P@k, MRR	Answer List
P48	2014	Question Classification, Candidate Answers Ranking	BOLT, TAC	English, Chinese, Arabic	Opinion	Open-Domain	F1	Answer List
P49	2014	Candidate Answers Extraction, Candidate Answers Ranking	Created by Authors	Korean	Definition	Open-Domain	F1, P@k, MRR, Recall@K	Web
P50	2013	Candidate Answers Ranking	NTCIR-6 QAC	Japonese	Why	Open-Domain	P@1, MAP	Documents
P51	2013	Question Classification	Created by Authors	Arabic	Opinion	Political	Precision	Documents
P52	2013	Question Classification	Yahoo! Answers	English	Comparison	Open-Domain	F1	-
P53	2013	Candidate Answers Ranking	Yahoo! Answers	English	How	Open-Domain	MRR	Answer List
P54	2013	Candidate Answers Ranking	ResPubliQA (CLEF 2010)	English, Italian	-	Open-Domain	Accuracy, MRR	Documents
P55	2012	Candidate Answers Extraction, Candidate Answers Ranking	Yahoo! Answers, Created by Authorts, Yahoo! Chiebukuro	Japonese	Why	Open-Domain	P@k, MAP	Documents, Web
P56	2012	Candidate Answers Extraction	Created by Authors	Japonese	How	Open-Domain	F1	Web
P57	2012	Candidate Answers Extraction	Created by Authors	Arabic	Why, How	Open-Domain	Accuracy	Web
P58	2012	Question Classification, Question Reformulation, Candidate Answers Ranking	Created by Authors	English	-	Open-Domain	F1, Rouge	Knowledge Graph
P59	2012	Candidate Answers Extraction	Yahoo! Answers	Japonese	Why	Open-Domain	F1, Accuracy	Web
P60	2011	Question Focus Recognition, Question Reformulation, Passage Extraction	Clinical Questions Collection	English	-	Health	Human Qualitative	Documents
P61	2011	Question Classification	Created by Authors	English	Why, How, Definition, Confirmation, Factoid	Open-Domain	P@1, Binary Precision	Documents
P62	2011	Question Classification	Jeopardy!	English	Definition, Factoid	Open-Domain	F1	Documents, Knowledge Graph
P63	2011	Candidate Answers Extraction	Yahoo! Answers	English	-	Open-Domain	Rouge	Answer List
P64	2011	Question Reformulation, Candidate Answers Ranking	Yahoo! Answers	English	How	Open-Domain	P@k, MRR	Answer List
P65	2011	Candidate Answers Extraction, Candidate Answers Ranking	Created by Authors	Chinese	Why, How	Open-Domain	P@k, MRR	Documents
P66	2011	Candidate Answers Extraction, Candidate Answers Ranking	NTCIR-8 CCLQA	Chinese	Why, Definition	Open-Domain	F1	Documents, Web
P67	2011	Candidate Answers Extraction, Candidate Answers Ranking	Yahoo! Answers	English	How	Open-Domain	P@k, MRR, Recall@K	Web, Answer List
P68	2010	Candidate Answers Ranking	WEB-QA, TREC-QA	English	Why, How, Definition	Open-Domain	F1	Documents, Web
P69	2010	Candidate Answers Extraction	Yahoo! Answers	English	Why	Open-Domain	F1	Documents
P70	2010	Candidate Answers Ranking	Yahoo! Answers, TREC-QA	English	Definition, Definition, Opinion	Open-Domain	F1, pyramid F-score	Documents, Web
P71	2010	Question Reformulation	MPQA	English	Opinion	Open-Domain	F1	Documents, Web
P72	2010	Candidate Answers Extraction, Candidate Answers Ranking	Created by Authors	Arabic	Definition	Open-Domain	P@k	Documents, Web
P73	2010	Candidate Answers Ranking	Created by Authors	Chinese	-	Open-Domain	P@k, MRR	Web
P74	2010	Question Classification, Question Reformulation	Created by Authors	English	-	Health	F1	Documents
P75	2010	Candidate Answers Ranking	TREC-QA, Created by Authors	English	Definition	Open-Domain	F1, MAP	Web
P76	2021	Candidate Answer Ranking	EPD, CCD	Chinese	-	Environment Protection, Childcare	MRR, P@K	Documents, Web
P77	2022	Candidate Answers Extraction, Reasoning	WN18RN, FB15k-237, SimpleQuestions, WebQuestionsSP	English	Factoid, Non-Factoid	Open-Domain	MRR, Hits@10, F1 score	Document, Knowledge Graph
P78	2021	Document Retrieval, Answer Generation	CLEF-IP 2011 benchmark dataset and a real-world dataset obtained from Google patent repository	English	Factoid, Definition	Patents	ROUGE-L and METEOR	Documents
P79	2021	Passage Extraction	WikiPassageQA, ANTIQUE	English	-	Open-Domain	MRR, MAP, P@10	Knowledge Graph
P80	2021	Candidate Answer Ranking	SemEval 2015, SemEval 2017	English	Factoid	Open-Domain	F1, Accuracy, MAP	Answer List
P81	2021	Question Classification	TREC, CLEF, Moroccan school books	Arabic	Factoid	Open-Domain	Accuracy, Precision, Recall, F1	-
P82	2022	Candidate Answer Ranking	Yahoo! Answers, Quora, Answers.com	English	Why	Open-Domain	MRR	Documents
P83	2022	Candidate Answers Extraction, Reasoning	Risk Models, CE Pairs, NATO-SDA, SemEval, Twitter	English	Confirmation	Open-Domain	Accuracy, Precision, Recall, F1	Knowledge Graph
P84	2022	Passage Ranking	WikiPassageQA, WikiQA, MS MARCO	English	Factoid	Open-Domain	MAP, MRR	Documents
P85	2021	Candidate Answer Ranking	QuoraQA, AmazonQA, YahooQA	English	-	Open-Domain	MRR and P@1	Web
P86	2022	Candidate Answer Extraction, Candidate Answer Ranking	WikiPassageQA, TREC-QA	English	-	Open-Domain	MAP, MRR	Documents
P87	2023	Question Classification, Candidate Answer Extraction, Answer Generation	CQA dataset from Stack Exchange (Created)	English	-	Open-Domain	Manual	Documents
P88	2022	Answer Generation	Persian religious dataset	Persian	-	Religious	ROUGE, BLEU	None
P89	2022	Passage Extraction	Arabic WikiReading, KaifLematha	Arabic	Factoid	Open-Domain	Exact match, F1	Document, Knowledge Graph

ID	Year	Tasks	Datasets	Language	Question Type	Domain of Knowledge	Evaluation Metrics	Knowledge Source Type
P90	2021	Candidate Answer Ranking	WikiQA, TREC QA, InsuranceQA, and Yahoo QA	English	Factoid	Open-Domain, Insurance	MAP, MRR	Knowledge Graph
P91	2021	Question Classification, Candidate Answer Extraction	allrecipes.com	English	Factoid, How	Cooking	Manual	Documents
P92	2021	Answer Generation	Tokyo Metropolitan Assembly	Japanese	-	Tokyo Metropolitan Assembly	ROUGE	Documents
P93	2021	Question Reformulation, Document Retrieval	Student Essays, Web Discourse	English	Comparison	Educational	nDCG	Documents
P94	2022	Question Classification	Created	English	Comparison	Open-Domain	F1	-
P95	2022	Question Classification	TREC, CLEF, Moroccan school books	Arabic	Factoid	Open-Domain, Educational	Accuracy, Recall, Precision, F1	-
P96	2022	Question Classification, Candidate Answer Ranking	TREC-QA and Wiki-QA	English	Factoid	Open-Domain	MAP, MRR	Documents
P97	2021	Candidate Answer Extraction	Yahoo! Chiebukuro	Japanese	How	Open-Domain	Exact Match, Partial Match	Documents
P98	2023	Candidate Answer Extraction	Smithsonian American Art Museum (SAAM) [Created]	English	Comparison, Confirmation, Factoid	Cultural Heritage	Precision, Recall, F1 score	Knowledge Graphs
P99	2022	Answer Generation	QuAC	English	-	Open-Domain	F1, Manual	Documents
P100	2022	Question Classification	SELECTED_SUBSET_ FROM(Yahoo nfL6, TREC 2007, Wikipidia Question-Answer Dataset)		Definition, Confirmation, Factoid	Open-Domain	Precision, Recall, F1	-
P101	2023	Answer Generation	CommonsenseQA, PiQA, HotpotQA	English	Factoid	Open-Domain, Physical Interaction	Accuracy	None
P102	2022	Question Classification	Yahoo Non-Factoid Question, TREC 2007, Wikipedia	English	Factoid, Definition, Confirmation	Open-Domain	Accuracy, Precision, Recall, F1	-
P103	2023	Candidate Answer Ranking, Answer Generation	HealthQA, NFCorpus	English	-	Health	MAP, NDCG, P@K, Recall@k	Documents, Knowledge Graphs
P104	2022	Passage Extraction, Candidate Answer Extraction, Candidate Answer Ranking	CORD-19	English	Factoid	Health	NDCG, Recall@K	Documents
P105	2023	Answer Generation	Wikihow Japonese (Created)	Japanese	How	Open-Domain	BLEU, ROUGE	Documents
P106	2023	Candidate Answer Extraction, Candidate Answer Ranking	Colected from web. (hellosehat.com, alodokter.com, and halodoc.com)	Indonesian	Definition, Why	Health	MRR, MAP	Documents
P107	2021	Candidate Answer Extraction, Candidate Answer Ranking	SemEval-2016 CQA	English	Factoid	Open-Domain	MRR, MAP	Documents
P108	2021	Candidate Answer Extraction, Candidate Answer Ranking	NLQuAD (Created)	English	-	Open-Domain	Exact match, Precision, Recall, F1	Documents
P109	2021	Candidate Answer Extraction, Candidate Answer Ranking	Wikipedia (Created)	Indonesian	Definition, Why	Science	MAP, MRR	Documents
P110	2023	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	WikiHow, PubMedQA	English	-	Health	ROUGE	Documents
P110	2021	Candidate Answer Extraction, Candidate Answer Ranking	NFPassageQA (Created)	English	-	Open-Domain	Precision@k, Recall@k, NDCG@k, Precision-IA@k, S-Recall@k	Documents
P112	2023	Candidate Answer Extraction, Candidate Answer Ranking	PQuAD	Persian	Factoid	Open-Domain	Exact match, F1	Documents
P113	2023	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	QASA (Created)	English	-	Computer Science	Precision, Recall, F1, ROUGE	Documents
P114	2022	Question Reformulation, Document Retrieval, Candidate Answer Ranking	Student Essays, Web Discourse	English	Comparison	Open-Domain	nDCG@5, Accuracy, Precision, Recall, F1	Documents
P115	2021	Candidate Answer Ranking	Qatar Living Forum	English	-	Open-Domain	MAP, MRR, Precision, Recall, F1, and Accuracy	Answer List
P116	2022	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	Constructed from the book "Determination and Prevention and Control of Food Safety Accidents.	Chinese	-	Food Safety	EM, F1, Precision, Recall	Knowledge Graph
P117	2021	Candidate Answer Ranking	120ASK	Chinese	-	Health	Accuracy, Precision, Recall, F1-score	Documents
P118	2023	Candidate Answer Extraction	So2al-wa-Gwab, Arabic SQuAD, ARCD, TyDi QA, AQAD, MLQA, AAQAD	Arabic	Factoid	Open-Domain	EM, F1-Score	Documents
P119	2022	Candidate Answer Extraction, Candidate Answer Ranking	QALD-5, QALD-9, LC-QuAD, ComplexQuestions	English	-	Open-Domain	Precision, Recall, F1-score	Knowledge Graph
P120	2022	Aggregate question answering, Candidate Answer Ranking	LC-QuAD, ComplexWebQuestion , QALD	English	-	Open-Domain	Precision, Recall, F1, Hit@1	Knowledge Graph
P121	2023	Question Classification, Candidate Answer Ranking	Yahoo! Webscope L-31, Yahoo! Answers User Profiles	English	-	Open-Domain	F1, MRR, Accuracy	Answer List
P122	2022	Passage Extraction, Candidate Answer Extraction, Answer Generation	Extracted Sample (MS MARCO, Google Natural Questions)	English	Comparison	Open-Domain	Precision, Recall, F1, Accuracy	Documents
P123	2021	Document Retrieval, Candidate Answer Ranking	PRIVACYQA	English	-	Law	Precision, Recall, F1, MRR	Documents
P124	2022	Document Retrieval, Candidate Answer Ranking	WDRASS	English	Factoid	Open-Domain	P@k, H@k	Documents

ID	Year	Tasks	Datasets	Language	Question Type	Domain of Knowledge	Evaluation Metrics	Knowledge Source Type
P125	2023	Answer Generation	WikiHowQA	English	How	Open-Domain	ROUGE, BERTScore, Manual	Documents

Appendix B: Systematic Review - References

ID	Paper
P1	Hashemi, Helia, et al. "ANTIQUE: A non-factoid question answering benchmark." Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42. Springer International Publishing, 2020.
P2	Dimitrakis, Eleftherios, et al. "Enabling efficient question answering over hundreds of linked datasets." Information Search, Integration, and Personalization: 13th International Workshop, ISIP 2019, Heraklion, Greece, May 9–10, 2019, Revised Selected Papers 13. Springer International Publishing, 2020.
P3	Bondarenko, Alexander, et al. "Comparative web search questions." Proceedings of the 13th International Conference on Web Search and Data Mining. 2020.
P4	Sarrouti, Mourad, and Said Ouatik El Alaoui. "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions." Artificial intelligence in medicine 102 (2020): 101767.
P5	Chen, Tengyang, et al. "Developing a how-to tip machine comprehension dataset and its evaluation in machine comprehension by BERT." Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER). 2020.
P6	Srinath, Mukund. "Convolutional Neural Network and Question Generation Based Approaches to Select Best Answers for Non-Factoid Questions." (2019).
P7	Zhu, Ming, et al. "A hierarchical attention retrieval model for healthcare question answering." The World Wide Web Conference. 2019.
P8	Li, Xuelian, et al. "A hybrid framework for problem solving of comparative questions." IEEE Access 7 (2019): 185961-185976.
P9	Yang, Min, et al. "Advanced community question answering by leveraging external knowledge and multi-task learning." Knowledge-Based Systems 171 (2019): 106-119.
P10	Ye, Qiaofei, et al. "A sentiment based non-factoid question-answering framework." 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019.
P11	Zhu, Nana, Zhijun Zhang, and Haiqun Ma. "Ranking answers of comparative questions using heterogeneous information organization from social media." Signal, Image and Video Processing 13 (2019): 1267-1274.
P12	Kulkarni, Ashish, et al. "Productqna: Answering user questions on e-commerce product pages." Companion proceedings of the 2019 world wide web conference. 2019.
P13	Cohen, Daniel, Liu Yang, and W. Bruce Croft. "WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval." The 41st international ACM SIGIR conference on research & development in information retrieval. 2018.
P14	Yulianti, Evi, et al. "Document summarization for answering non-factoid queries." IEEE transactions on knowledge and data engineering 30.1 (2017): 15-28.
P15	Costa, Jose Ortiz, and Anagha Kulkarni. "Leveraging knowledge graph for open-domain question answering." 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). IEEE, 2018.
P16	Ha, Thi-Thanh, et al. "Unsupervised Sentence Embeddings for Answer Summarization in Non-factoid CQA." Computación y Sistemas 22.3 (2018): 835-843.
P17	Tran, Nam Khanh, and Claudia Niederée. "A neural network-based framework for non-factoid question answering." Companion Proceedings of the The Web Conference 2018. 2018.
P18	Tran, Nam Khanh, and Claudia Niederée. "A neural network-based framework for non-factoid question answering." Companion Proceedings of the The Web Conference 2018. 2018.
P19	Pithyaachariyakul, Chanin, and Anagha Kulkarni. "Automated question answering system for community-based questions." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
P20	Herrera, Jose, Barbara Poblete, and Denis Parra. "Learning to leverage microblog information for QA retrieval." European Conference on Information Retrieval. Cham: Springer International Publishing, 2018.
P21	Vikraman, Lakshmi, W. Bruce Croft, and Brendan O'Connor. "Exploring diversification in non-factoid question answering." Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval. 2018.
P22	Sharma, Akshay, and Chetan Harithas. "Inner attention based bi-lstms with indexing for non-factoid question answering." 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2018.
P23	Yulianti, Evi, et al. "Ranking documents by answer-passage quality." The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018.
P24	Khvalchik, Maria, and Anagha Kulkarni. "Open-domain non-factoid question answering." Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20. Springer International Publishing, 2017.
p25	Wang, Di, and Eric Nyberg. "CMU OAQA at TREC 2016 LiveQA: An Attentional Neural Encoder-Decoder Approach for Answer Ranking." TREC. 2016.
P26	Gao, Xiang, Kai Niu, and Zhiqiang He. "A convolutional neural network model for non-factoid Chinese answer selection." 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). IEEE, 2017.
P27	Ma, Rongqiang, et al. "Hybrid answer selection model for non-factoid question answering." 2017 international conference on asian language processing (IALP). IEEE, 2017.
P28	Khvalchik, Maria, Chanin Pithyaachariyakul, and Anagha Kulkarni. "Answering the Hard Questions." Language, Data, and Knowledge: First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings 1. Springer International Publishing, 2017.
P29	Song, Hongya, et al. "Summarizing answers in non-factoid community question-answering." Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017.
P30	Jones, Gareth JF, et al., eds. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings. Vol. 10456. Springer, 2017.
P31	Cohen, Daniel, and W. Bruce Croft. "End to end long short term memory networks for non-factoid question answering." Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. 2016.
P32	Yang, Liu, et al. "Beyond factoid QA: effective methods for non-factoid answer sentence retrieval." Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38. Springer International Publishing, 2016.
P33	Khodadi, Iman, and Mohammad Saniee Abadeh. "Genetic programming-based feature learning for question answering." Information Processing & Management 52.2 (2016): 340-357.
P34	Saelan, A., A. Purwarianti, and D. H. Widyantoro. "Question analysis for Indonesian comparative question." Journal of Physics: Conference Series. Vol. 801. No. 1. IOP Publishing, 2017.
P35	Azmi, Aqil M., and Nouf A. Alshenaifi. "Answering arabic why-questions: Baseline vs. rst-based approach." ACM Transactions on Information Systems (TOIS) 35.1 (2016): 1-19.
P36	Bao, Xin-Qi, and Yun-Fang Wu. "A tensor neural network with layerwise pretraining: Towards effective answer retrieval." Journal of Computer Science and Technology 31.6 (2016): 1151-1160.
P37	Yulianti, Evi, et al. "Using semantic and context features for answer summary extraction." Proceedings of the 21st Australasian Document Computing Symposium. 2016.

ID	Paper
P38	Tymoshenko, Kateryna, Daniele Bonadiman, and Alessandro Moschitti. "Learning to rank non-factoid answers: Comment selection in web forums." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016.
P39	Pang, Liang, et al. "SPAN: understanding a question with its support answers." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 30. No. 1. 2016.
P40	Wang, Bingning, et al. "Employing External Rich Knowledge for Machine Comprehension." IJCAI. 2016.
P41	Yu, Yang, et al. "End-to-end answer chunk extraction and ranking for reading comprehension." arXiv preprint arXiv:1610.09996 (2016).
P42	Wu, Youzheng, et al. "Leveraging social Q&A collections for improving complex question answering." Computer Speech & Language 29.1 (2015): 1-19.
P43	Feng, Minwei, et al. "Applying deep learning to answer selection: A study and an open task." 2015 IEEE workshop on automatic speech recognition and understanding (ASRU). IEEE, 2015.
P44	Xie, Zongsheng, et al. "Answer quality assessment in CQA based on similar support sets." Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, Guangzhou, China, November 13-14, 2015, Proceedings 14. Springer International Publishing, 2015.
P45	Tymoshenko, Kateryna, and Alessandro Moschitti. "Assessing the impact of syntactic and semantic structures for answer passages reranking." Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015.
P46	Chen, Ruey-Cheng, et al. "Harnessing semantics for answer sentence retrieval." Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval. 2015.
P47	Jansen, Peter, Mihai Surdeanu, and Peter Clark. "Discourse complements lexical semantics for non-factoid answer reranking." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.
P48	Chaturvedi, Snigdha, et al. "Joint question clustering and relevance prediction for open domain non-factoid question answering." Proceedings of the 23rd international conference on World wide web. 2014.
P49	Ryu, Pum-Mo, Myung-Gil Jang, and Hyun-Ki Kim. "Open domain question answering using Wikipedia-based knowledge model." Information Processing & Management 50.5 (2014): 683-692.
P50	Oh, Jong-Hoon, et al. "Why-question answering using intra-and inter-sentential causal relations." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013.
P51	Bayoudhi, Amine, Hatem Ghorbel, and Lamia Hadrich Belguith. "Question answering system for dialogues: A new taxonomy of opinion questions." Flexible Query Answering Systems: 10th International Conference, FQAS 2013, Granada, Spain, September 18-20, 2013. Proceedings 10. Springer Berlin Heidelberg, 2013.
P52	Li, Shasha, et al. "Comparable entity mining from comparative questions." IEEE transactions on knowledge and data engineering 25.7 (2011): 1498-1509.
P53	Atkinson, John, Alejandro Figueroa, and Christian Andrade. "Evolutionary optimization for ranking how-to questions based on user-generated contents." Expert Systems with Applications 40.17 (2013): 7060-7068.
P54	Molino, Piero. "Semantic models for answer re-ranking in question answering." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. 2013.
P55	Oh, Jong-Hoon, et al. "Why question answering using sentiment analysis and word classes." Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. 2012.
P56	Ishikawa, Kyohei, and Hayato Ohwada. "Extraction of how-to type question-answering sentences using query sets." Knowledge Management and Acquisition for Intelligent Systems: 12th Pacific Rim Knowledge Acquisition Workshop, PKAW 2012, Kuching, Malaysia, September 5-6, 2012. Proceedings 12. Springer Berlin Heidelberg, 2012.
P57	Sadek, Jawad, Fairouz Chakkour, and Farid Meziane. "Arabic rhetorical relations extraction for answering" why" and" how to" questions." Natural Language Processing and Information Systems: 17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26-28, 2012. Proceedings 17. Springer Berlin Heidelberg, 2012.
P58	Yu, Jianxing, Zheng-Jun Zha, and Tat-Seng Chua. "Answering opinion questions on products by exploiting hierarchical organization of consumer reviews." Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. 2012.
P59	Tanaka, Katsuyuki, Tetsuya Takiguchi, and Yasuo Ariki. "Towards domain independent why text segment classification based on bag of function words." AI 2012: Advances in Artificial Intelligence: 25th Australasian Joint Conference, Sydney, Australia, December 4-7, 2012. Proceedings 25. Springer Berlin Heidelberg, 2012.
P60	Cao, YongGang, et al. "AskHERMES: An online question answering system for complex clinical questions." Journal of biomedical informatics 44.2 (2011): 277-288.
P61	Fan, Shixi, et al. "Using hybrid kernel method for question classification in CQA." Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part III 18. Springer Berlin Heidelberg, 2011.
P62	Moschitti, Alessandro, et al. "Using syntactic and semantic structural kernels for classifying definition questions in Jeopardy!." Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011.
P63	Liu, Xiaoying, et al. "Using concept-level random walk model and global inference algorithm for answer summarization." Information Retrieval Technology: 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings 7. Springer Berlin Heidelberg, 2011.
P64	Hieber, Felix, and Stefan Riezler. "Improved answer ranking in social question-answering portals." Proceedings of the 3rd international workshop on Search and mining user-generated contents. 2011.
P65	Zong, Huanyun, et al. "An answer extraction method based on discourse structure and rank learning." 2011 7th International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2011.
P66	Ren, Han, et al. "A Web knowledge based approach for complex question answering." Information Retrieval Technology: 7th Asia Information Retrieval Societies Conference, AIRS 2011, Dubai, United Arab Emirates, December 18-20, 2011. Proceedings 7. Springer Berlin Heidelberg, 2011.
P67	Surdeanu, Mihai, Massimiliano Ciaramita, and Hugo Zaragoza. "Learning to rank answers to non-factoid questions from web collections." Computational linguistics 37.2 (2011): 351-383.
P68	Quarteroni, Silvia, and Alessandro Moschitti. "A Comprehensive Resource to Evaluate Complex Open Domain Question Answering." LREC. 2010.
P69	Nagy, Iulia, Katsuyuki Tanaka, and Yasuo Ariki. "Why text segment classification based on part of speech feature selection." Discovery Science: 13th International Conference, DS 2010, Canberra, Australia, October 6-8, 2010. Proceedings 13. Springer Berlin Heidelberg, 2010.
P70	Achananuparp, Palakorn, et al. "Answer diversification for complex question answering on the web." Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14. Springer Berlin Heidelberg, 2010.

ID	Paper
P71	Miao, Yajie, and Chunping Li. "Mining wikipedia and yahoo! answers for question expansion in opinion qa." Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14. Springer Berlin Heidelberg, 2010.
P72	Trigui, Omar, Lamia Hadrich Belguith, and Paolo Rosso. "An automatic definition extraction in Arabic language." Natural Language Processing and Information Systems: 15th International Conference on Applications of Natural Language to Information Systems, NLDB 2010, Cardiff, UK, June 23-25, 2010. Proceedings 15. Springer Berlin Heidelberg, 2010.
P73	Wang, Baoxun, et al. "Modeling semantic relevance for question-answer pairs in web social communities." Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010.
P74	Cao, Yong-gang, et al. "Automatically extracting information needs from complex clinical questions." Journal of biomedical informatics 43.6 (2010): 962-971.
P75	Figueroa, Alejandro, and John Atkinson. "Answering definition questions: Dealing with data sparseness in lexicalised dependency trees-based language models." Web Information Systems and Technologies: 5th International Conference, WEBIST 2009, Lisbon, Portugal, March 23-26, 2009, Revised Selected Papers 5. Springer Berlin Heidelberg, 2010.
P76	Lv, Ming-Qi, et al. "Domain-Specific Non-Factoid Question Answering System based on Terminology Mining and Siamese Neural Network." Journal of Information Science & Engineering 37.4 (2021).
P77	Xu, Sa. "A Knowledge Reasoning Model Based on Non-Factoid Information Enhancement." Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City. 2022.
P78	Zihayat, Morteza, and Rochelle Etwaroo. "A non-factoid question answering system for prior art search." Expert Systems with Applications 177 (2021): 114910.
P79	TR, Akila Devi, et al. "A novel framework using zero shot learning technique for a non-factoid question answering system." International Journal of Web-Based Learning and Teaching Technologies (IJWLTT) 16.6 (2021): 1-13.
P80	Yang, Haitian, et al. "Amqan: Adaptive multi-attention question-answer networks for answer selection." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III. Springer International Publishing, 2021.
P81	Hamza, Alami, et al. "An arabic question classification method based on new taxonomy and continuous distributed representation of words." Journal of King Saud University-Computer and Information Sciences 33.2 (2021): 218-224.
P82	Breja, Manvi, and Sanjay Kumar Jain. "Analyzing linguistic features for answer re-ranking of why-questions." Journal of Cases on Information Technology (JCIT) 24.3 (2022): 1-16.
P83	Kayesh, Humayun, Md Saiful Islam, and Junhu Wang. "Answering binary causal questions using role-oriented concept embedding." IEEE Transactions on Artificial Intelligence (2022).
P84	Lin, Dengwen, et al. "BERT-SMAP: Paying attention to Essential Terms in passage ranking beyond BERT." Information Processing & Management 59.2 (2022): 102788.
P85	Su, Lixin, et al. "Beyond relevance: Trustworthy answer selection via consensus verification." Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021.
P86	Jin, Zhiling, et al. "Bi-granularity Adversarial Training for Non-factoid Answer Retrieval." European Conference on Information Retrieval. Cham: Springer International Publishing, 2022.
P87	Jin, Sol, et al. "Building a deep learning-based QA system from a CQA dataset." Decision Support Systems 175 (2023): 114038.
P88	Bidgoly, Amir Jalaly, Hossein Amirkhani, and Razieh Baradaran. "Clustering-based Sequence to Sequence Model for Generative Question Answering in a Low-resource Language." ACM Transactions on Asian and Low-Resource Language Information Processing 22.2 (2022): 1-14.
P89	Albilali, Eman, Nora Al-Twairesh, and Manar Hosny. "Constructing arabic reading comprehension datasets: Arabic wikireading and kaiflematha." Language Resources and Evaluation 56.3 (2022): 729-764.
P90	Deng, Yang, et al. "Contextualized knowledge-aware attentive neural network: Enhancing answer selection with knowledge." ACM Transactions on Information Systems (TOIS) 40.1 (2021): 1-33.
P91	Khilji, Abdullah Faiz Ur Rahman, et al. "CookingQA: answering questions and recommending recipes based on ingredients." Arabian Journal for Science and Engineering 46.4 (2021): 3701-3712.
P92	Kawai, Teruya, Tomoyosi Akiba, and Shigeru Masuyama. "Development of Political QA Systems aimed at Assembly Minutes based on Abstractive Summarization." 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA). IEEE, 2021.
P93	Alhamzeh, Alaa, et al. "Distilbert-based argumentation retrieval for answering comparative questions." Conference and Labs of the Evaluation Forum (CLEF 2021). Vol. 2936. 2021.
P94	Beloucif, Meriem, et al. "Elvis vs. M. Jackson: Who has more albums? classification and identification of elements in comparative questions." 13th International Conference on Language Resources and Evaluation (LREC), JUN 20-25, 2022, Marseille, FRANCE. European Language Resources Association, 2022.
P95	Hamza, Alami, et al. "Embedding arabic questions by feature-level fusion of word representations for questions classification: It is worth doing?." Journal of King Saud University-Computer and Information Sciences 34.9 (2022): 6583-6594.
P96	Abbasiantaeb, Zahra, and Saeedeh Momtazi. "Entity-aware answer sentence selection for question answering with transformer-based language models." Journal of Intelligent Information Systems 59.3 (2022): 755-777.
P97	Li, Tingxuan, Shuting Bai, and Takehito Utsuro Fuzhu Zhu. "Evaluating a How-to Tip Machine Comprehension Model with QA Examples collected from a Community QA Site." Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. 2021.
P98	Gounakis, Nikos, Michalis Mountantonakis, and Yannis Tzitzikas. "Evaluating a radius-based pipeline for question answering over cultural (CIDOC-CRM based) knowledge graphs." Proceedings of the 34th ACM Conference on Hypertext and Social Media. 2023.
P99	Zhang, Zhiyuan, Qiaoqiao Feng, and Yujie Wang. "Explicit History Selection for Conversational Question Answering." 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2022.
P100	Mohasseb, Alaa, and Andreas Kanavos. "Factoid vs. non-factoid question identification: An ensemble learning approach." 18th International Conference on Web Information Systems and Technologies. SciTePress, 2022.
P101	Espejel, Jessica López, et al. "GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts." Natural Language Processing Journal 5 (2023): 100032.
P102	Mohasseb, Alaa, and Andreas Kanavos. "Grammar-Based Question Classification Using Ensemble Learning Algorithms." International Conference on Web Information Systems and Technologies. Cham: Springer Nature Switzerland, 2022.
P103	Wang, Xiaoli, et al. "How context or knowledge can benefit healthcare question answering?." IEEE Transactions on Knowledge and Data Engineering 35.1 (2021): 575-588.
P104	Otegi, Arantxa, et al. "Information retrieval and question answering: A case study on COVID-19 scientific literature." Knowledge-Based Systems 240 (2022): 108072.
P105	Wang, Xiaotian, et al. "Japanese how-to tip machine reading comprehension by multi-task learning based on generative model." International Conference on Text, Speech, and Dialogue. Cham: Springer Nature Switzerland, 2023.

ID	Paper
P106	Kusumaningrum, Retno, et al. "Long Short-Term Memory for Non-Factoid Answer Selection in Indonesian Question Answering System for Health Information." International Journal of Advanced Computer Science and Applications 14.2 (2023).
P107	Zhang, Yingxue, et al. "MS-Ranker: Accumulating evidence from potentially correct candidates via reinforcement learning for answer selection." Neurocomputing 449 (2021): 270-279.
P108	Soleimani, Amir, Christof Monz, and Marcel Worring. "NLQuAD: A non-factoid long question answering data set." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021.
P109	Hanifah, Alfi Fauzia, and Retno Kusumaningrum. "Non-factoid answer selection in indonesian science question answering system using long short-term memory (LSTM)." Procedia Computer Science 179 (2021): 736-746.
P110	Deng, Yang, et al. "Nonfactoid question answering as query-focused summarization with graph-enhanced multihop inference." IEEE Transactions on Neural Networks and Learning Systems (2023).
P110	Vikraman, Lakshmi, et al. "Passage similarity and diversification in non-factoid question answering." Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 2021.
P112	Darvishi, Kasra, et al. "PQuAD: A Persian question answering dataset." Computer Speech & Language 80 (2023): 101486.
P113	Lee, Yoonjoo, et al. "QASA: advanced question answering on scientific articles." International Conference on Machine Learning. PMLR, 2023.
P114	Alhamzeh, Alaa, et al. "Query Expansion, Argument Mining and Document Scoring for an Efficient Question Answering System." International Conference of the Cross-Language Evaluation Forum for European Languages. Cham: Springer International Publishing, 2022.
P115	Arif, Rehab, and Maryam Bashir. "Question Answer Re-Ranking using Syntactic Relationship." 2021 15th International Conference on Open Source Systems and Technologies (ICOSST). IEEE, 2021.
P116	Shi, Yuntao, et al. "Research on food safety multi-hop reasoning question answering based on cognitive graph." 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2022.
P117	Lin, Cheng Ying, Yi-Hung Wu, and Arbee LP Chen. "Selecting the most helpful answers in online health question answering communities." Journal of Intelligent Information Systems 57.2 (2021): 271-293.
P118	Al-Omari, Hani, and Rehab Duwairi. "So2al-wa-Gwab: A New Arabic Question-Answering Dataset Trained on Answer Extraction Models." ACM Transactions on Asian and Low-Resource Language Information Processing 22.8 (2023): 1-21.
P119	Bakhshi, Mahdi, et al. "SParseQA: Sequential word reordering and parsing for answering complex natural language questions over knowledge graphs." Knowledge-Based Systems 235 (2022): 107626.
P120	Wu, Shaojuan, et al. "Structure-sensitive semantic matching for aggregate question answering over knowledge base." Journal of Web Semantics 74 (2022): 100737.
P121	Trewhela, Alvaro, and Alejandro Figueroa. "Text-based neural networks for question intent recognition." Engineering Applications of Artificial Intelligence 121 (2023): 105933.
P122	Bondarenko, Alexander, et al. "Towards understanding and answering comparative questions." Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022.
P123	Vold, Andrew, and Jack G. Conrad. "Using transformers to improve answer retrieval for legal questions." Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. 2021.
P124	Zhang, Zeyu, et al. "Wdrass: A web-scale dataset for document retrieval and answer sentence selection." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
P125	Bolotova-Baranova, Valeriia, et al. "WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.

Appendix C: Resumo expandido

Os sistemas de Question Answering (QA) são projetados para responder perguntas usando a linguagem natural, fornecendo respostas precisas e informativas. Perguntas que demandam respostas mais detalhadas, como "Por que o céu é azul?", acrescentam uma complexidade adicional a esses sistemas. Além disso, avaliar as respostas extensas desses sistemas é uma tarefa desafiadora, pois quanto mais longo o texto a ser avaliado, maior são as possibilidades de expressar essa informação. Isso complica a comparação com uma resposta de referência, que pode ser semanticamente idêntica, mas estruturalmente diferente.

Métricas automáticas de avaliação de texto que quantificam a similaridade entre a resposta fornecida e uma resposta de referência, visam sintetizar múltiplas características, como precisão, completude, relevância e fluência, em uma única medida, o que pode dificultar a avaliação adequada de cada característica distinta.

Este trabalho propõe uma análise especificamente voltada para os critérios de completude e relevância em respostas longas geradas por sistemas de QA. Busca-se desenvolver uma metodologia para avaliar esses critérios, por meio da aplicação de métricas automáticas, permitindo um entendimento mais profundo e detalhado da eficácia dos sistemas de QA.

A pesquisa propôs primeiramente uma revisão sistemática de sistemas de QA não factóides, abordando as estratégias de avaliação utilizadas e identificando lacunas, especialmente na avaliação de respostas longas e detalhadas. Com base nesta revisão, é desenvolvido um conjunto de dados anotados especificamente para avaliar a completude e a relevância. Este conjunto de dados inclui 106 perguntas do subreddit "Explain Like I'm Five", cada uma com duas respostas: uma gerada por humanos e outra pelo modelo GPT-4. As respostas foram avaliadas por humanos utilizando uma ferramenta que permitia atribuir notas de 0 a 100 para cada critério.

Neste trabalho, foi proposto três modelos de métricas para avaliar completude e relevância:

- Estratégia baseada em prompts com Modelos de Linguagem de Grande Escala (LLMs): Utiliza prompts para orientar o modelo GPT-4 a analisar e pontuar as respostas.
- Modelo que adapta conceitos de precisão e revocação: Avalia a completude e a relevância segmentando a resposta em unidades discretas de informação.
- Modelo de regressão treinado com dados sintéticos: Atribui pontuações de completude e relevância com base em características sintetizadas das respostas.

Os resultados demonstraram que a estratégia de prompts com o GPT-4 se alinha bem com as avaliações humanas, especialmente em termos de completude, indicando uma forte correlação com o julgamento humano. O modelo de regressão também mostrou alta correlação na avaliação de completude, onde superou o GPT-4 em diferentes cenários. Além disso, essas métricas propostas foram comparadas com métricas convencionais, como BLEU, ROUGE e BERTScore, através de um benchmark que correlacionava as pontuações das métricas com as avaliações humanas.

As métricas desenvolvidas neste trabalho oferecem uma nova perspectiva na avaliação de respostas longas em sistemas de QA, focando em critérios específicos que são fundamentais para a utilidade das respostas. Além disso, os modelos propostos oferecem alternativas viáveis que não dependem de respostas de referência, possibilitando sua aplicação em contextos onde essas não estão disponíveis.

Esta pesquisa avança o estado da arte ao introduzir métodos que abordam as lacunas deixadas pelas abordagens tradicionais. Assim, apresenta avanços no campo de QA para respostas longas, propondo novas abordagens e técnicas de avaliação que se concentram especificamente nos critérios de completude e relevância. As principais contribuições deste trabalho são listadas a seguir:

- **Revisão Sistemática em QA Não-Factóide**: A revisão realizada oferece uma análise abrangente dos sistemas de QA não-factoidais, apresentando métodos, tarefas, conjuntos de dados, estratégias de avaliação e resultados obtidos. Essa análise destaca a complexidade das respostas mais extensas e a necessidade de métodos de avaliação. A revisão identifica lacunas na literatura existente, especialmente na capacidade dos sistemas de compor respostas detalhadas e considerar o contexto de várias fontes de informação. Assim, esta revisão sistemática não apenas resume o estado da arte em QA não-factoide, mas também estabelece um ponto de partida para futuros desenvolvimentos no campo, como visto neste trabalho.

- **Conjunto de Dados para Avaliação de Métricas**: Uma das principais contribuições deste trabalho é a criação de um conjunto de dados anotados focado em perguntas do tipo "Instrução" no campo da Ciência da Computação, onde as respostas são extensas e foram anotadas por humanos com base nos critérios de completude e relevância. Esse conjunto de dados pode ser usado como uma ferramenta para avaliar modelos métricos que focam nesses critérios, visando um ambiente controlado para minimizar vieses e ajudar na interpretabilidade dos resultados. Essa base de dados anotada pode ser usada como um recurso para testar e refinar métricas de avaliação. Também pode servir como base para pesquisas futuras destinadas a entender melhor como as respostas longas se relacionam com os critérios de completude e relevância.

- **Modelos Métricos Propostos para Avaliar Completude e Relevância**: Este trabalho contribui para o desenvolvimento de novos modelos métricos projetados especificamente para avaliar a completude e a relevância de respostas longas. As métricas visam entender como os sistemas de QA avaliados lidam com a profundidade e pertinência das informações fornecidas, abordando duas dimensões críticas que influenciam diretamente a utilidade das respostas aos usuários. Os modelos propostos incluem:
  - **Estratégia Baseada em Prompt com LLM Generativo**: Usando as capacidades avançadas de compreensão de texto dos LLMs, como o GPT-4, este modelo emprega uma técnica de prompt que orienta o LLM a analisar e pontuar a completude e relevância das respostas.
  - **Adaptação dos Conceitos de Revocação e Precisão**: Este modelo adapta métricas tradicionais de revocação e precisão para medir, respectivamente, a completude e relevância das respostas. Segmentando a resposta em unidades discretas de informação, o

modelo avalia quantitativamente quanto da informação relevante a resposta contém em relação ao que seria ideal (completude) e quanto do conteúdo da resposta é relevante em relação ao seu volume total (relevância).

    - **Modelo de Regressão Baseado em Dados Sintéticos**: Desenvolvido para prever pontuações de completude e relevância, este modelo é treinado com um conjunto de dados sintéticos que simula diferentes níveis de completude e relevância. Usando técnicas de Processamento de Linguagem Natural e modelos como o BERT para compreensão de texto, o modelo de regressão é capaz de atribuir valores numéricos às respostas, quantificando sua completude e relevância.

- **Avaliação e Comparação de Métricas para os Critérios de Completude e Relevância**: Esta pesquisa contribui para o campo de QA avaliando e comparando diferentes métricas de qualidade para respostas longas, com foco específico nos critérios de completude e relevância. A análise tanto das métricas convencionais quanto das métricas recém-propostas oferece insights sobre como cada uma se alinha com o julgamento humano, permitindo um entendimento das capacidades e limitações de cada método avaliativo. Por meio da aplicação de métricas convencionais como BLEU, ROUGE, BERTScore, entre outras, esta pesquisa determina até que ponto essas métricas são capazes de capturar os aspectos de completude e relevância das respostas.

Este estudo fornece uma base sólida para futuras pesquisas, visando a melhoria contínua dos sistemas de QA em fornecer respostas informativas e precisas, alinhadas às expectativas e necessidades dos usuários.