Journal of Computing and Security

Research Article, 2024, Volume 11, Number 2 (pp. 97–116)

http://www.jcomsec.org

Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models

Hossein Bahak [a], Farzaneh Taheri [a], Zahra Zojaji [a,*], Arefeh Kazemi [b]

[a] *Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran.*

[b] *ADAPT Research Centre, Dublin City University, Ireland.*

ARTICLE INFO.

Article history:

Received: 12 February 2024

Revised: 19 September 2024

Accepted: 28 September 2024

Published Online: 21 June 2025

Keywords:

ChatGPT, Question Answering Systems, Large Language Models, Performance Evaluation, Hallucination.

ABSTRACT

In the current era, a multitude of language models have emerged to cater to user inquiries. Notably, the GPT-3.5-Turbo language model has gained substantial attention as the underlying technology for ChatGPT. Leveraging extensive parameters, this model adeptly responds to a wide range of questions. However, due to its reliance on internal knowledge, the accuracy of responses may not be absolute. This article examines ChatGPT as a Question Answering System (QAS), comparing its performance to other existing QASs. The primary focus is on evaluating ChatGPT's efficiency in extracting responses from provided paragraphs, a core QAS capability. Additionally, performance comparisons are made in scenarios without a surrounding passage. Multiple experiments were conducted in ChatGPT, exploring response hallucination and considering question complexity. Evaluation employed well-known Question Answering (QA) datasets, including SQuAD, NewsQA, and PersianQuAD, across English and Persian languages. Metrics such as F score, exact match and accuracy were used in the assessment. The study reveals that, while ChatGPT demonstrates competence as a generative model, it is less effective in question answering compared to task-specific models. Providing context improves its performance, and prompt engineering improves precision, particularly for questions lacking explicit answers in the provided paragraphs. ChatGPT excels at simpler factual questions compared to the "how" and "why" question types. The evaluation highlights occurrences of hallucinations, where ChatGPT provides responses to questions without available answers in the provided context.

1 Introduction

The ability to answer questions accurately and effectively is a critical component of natural language understanding and communication systems. Question answering (QA) tasks have gained significant attention in the field of Artificial Intelligence (AI) as they represent a fundamental challenge for developing intelligent systems that can comprehend and respond to user queries in a human-like manner. With the advent of large scale language models, such as GPT 3.5 used by ChatGPT, which are based on advanced deep learning techniques, there is growing interest in evaluating their performance on QA tasks.

* Corresponding author.

Email address: z.zojaji@eng.ui.ac.ir (Z. Zojaji)

https://dx.doi.org/10.22108/jcs.2024.140680.1141

ISSN: 2322-4460

This paper aims to provide a comprehensive examination of ChatGPT's effectiveness on the task of question answering. Despite the limited research in this area, no studies have evaluated ChatGPT's ability to extract answers from given paragraphs. When used as a QAS without any text for answer extraction, ChatGPT must rely on its knowledge to generate answers, which can be strongly influenced by training bias. Additionally, the problem of hallucination arises when ChatGPT is unable to clearly identify the answer to a specific question. This study focuses on investigating ChatGPT's ability to extract answers from given paragraphs and comparing its performance to cases where no surrounding passage is provided for the answer. As the ability of a QAS to extract answers from related passages is key to its performance, this concentration is justified for analyzing ChatGPT's capabilities as a QAS.

We conducted a series of tests and evaluations using various question-answer datasets to assess the effectiveness and accuracy of ChatGPT in providing precise answers to user queries. By leveraging well-established benchmark datasets, such as the Stanford Question Answering Dataset (SQuAD), we can compare ChatGPT's performance against other state-of-the-art models and evaluate its capabilities in handling diverse question types and linguistic complexities. The choice of datasets is crucial for a comprehensive evaluation of ChatGPT's question-answer performance. In addition to SQuAD , we utilized NewsQA as a widely recognized QA dataset for news domain and and PersianQuAD, which covers a broad range of topics and question types in Persian language. By including multiple datasets, we ensure a robust evaluation that includes different domains, question styles, and linguistic variations.

To measure the performance of ChatGPT on question answering, we employ well-established metrics such as accuracy, F1 score, and Exact Match (EM). These metrics allow us to assess the model's ability to provide correct answers and accurately capture the span of the answer within the given context. By quantifying ChatGPT's performance using multiple metrics, we can gain a comprehensive understanding of its strengths and weaknesses in tackling the complexities of question answering tasks. Furthermore, we analyzed and compared ChatGPT's performance across different question categories, including factoid and nonfactoid questions. Factoid questions seek concise factual answers, while non-factoid questions require a deeper understanding of context and inference. By evaluating ChatGPT's performance on both types of question, we can evaluate its ability to handle a wide range of question styles and information retrieval challenges. In addition to evaluating ChatGPT's performance, we also conducted an in-depth analysis of its responses to gain insights into the model's decision-making process. By examining both correct and incorrect answers, we aim to uncover patterns, limitations, and potential areas for improvement in ChatGPT's question answering capabilities. This analysis provides valuable insights into the model's reasoning abilities and aids in understanding its strengths and limitations. The findings of this study have significant implications for the development and enhancement of question answering systems. Understanding the performance and limitations of ChatGPT on question answering tasks contributes to advancing the field of natural language understanding and facilitates the development of more accurate and reliable AI systems.

The key findings of the conducted research are summarized as follows.

- The performance of ChatGPT was compared with that of other models pretrained in various tasks, including question answering. The results indicated that ChatGPT, as a general generative model, performs less effectively than task-specific models in answering questions.
- It was observed that ChatGPT's performance significantly improves when provided with context for answering questions, as opposed to when no surrounding paragraph is given.
- Experimental findings suggest that prompt engineering greatly enhances ChatGPT's precision, particularly when the answer to a question is not present in the provided paragraph. Specifically, a two-step prompt querying the presence of the answer prior to requesting it yields superior performance compared to a one-step query.
- ChatGPT demonstrates proficiency in answering simpler, factual questions compared to "how" and "why" question types.
- Evaluation of hallucinations revealed that ChatGPT provides responses to a significant percentage of questions for which no answer is available in the provided context.

The findings of this study have significant implications for the development and enhancement of question answering systems. Understanding the performance and limitations of ChatGPT on question answering tasks contributes to advancing the field of natural language understanding and facilitates the development of more accurate and reliable AI systems.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of related work and highlight key advancements and challenges. Section 3 describes the evaluation framework used to assess the performance of ChatGPT. Section 4 presents the results of the evaluation of the ChatGPT's re-

sponses from different points of view. Finally, we conclude the paper in Section 5, summarizing the key insights and contributions of our study.

2 Related Works

Since the research is related to both QAS and ChatGPT evaluation studies, we analyzed the related work into three sections including evaluation of ChatGPT, review of QASs and the evaluation of ChatGPT performance on QASs.

2.1 Evaluating Various Capabilities of ChatGPT

There is limited research on the evaluation of ChatGPT as a QAS. However, a considerable research effort has been produced recently to evaluate various abilities of ChatGPT. Some general evaluation of ChatGPT was performed in several works (e.g. [1], [2],[3]). Some researchers [4], [5] investigated its mathematical problem solving ability. Zhong et al.[6] evaluated its understanding abilities. The performance of ChatGPT in fixing code bugs is examined by several works [7], [8], [9] and its overall applications in software engineering is studied in [10]. Again, it was also evaluated in translation ([11], [12], [13]) and summarization tasks ([14], [15]). Wang et al. [16] studied the out-of-distribution behaviors of ChatGPT. Several studies like [5] and [1] analyzed the reasoning ability of ChatGPT. The success of ChatGPT for open information extraction was also investigated by Li et al.[17].

2.2 Question Answering Systems

The field of Question Answering (QA) is widely recognized in Natural Language Processing (NLP), Information Extraction (IE), and Information Retrieval (IR). Initially, QA research emerged within the domain of IR. The first QA system, BaseBall [18], was developed in 1961 to answer specific questions about American baseball games. This system utilized punched cards to process questions and retrieve answers from stored data on baseball games, employing linguistic patterns to understand question meaning. Another notable system, LUNAR [19], was designed in 1973 to assist lunar geologists in finding answers related to NASA Apollo moon missions. This system leveraged information obtained from the missions to provide relevant answers.

The Text Retrieval Conference (TREC) played a significant role in advancing QA research. In 1999, TREC introduced a QA track, offering researchers a dataset comprising news articles, questions, and corresponding answers [20]. TREC continued to organize QA tracks in subsequent years, leading to significant advancements in the QA task.

With the rise of the World Wide Web and the proliferation of web pages, researchers started utilizing web documents as the primary information source for QA systems [21]. Modern QA systems are built upon web-based approaches, aiming to retrieve answers from online documents.

QA research has experienced notable progress in recent years, witnessing the development of numerous QA systems. Many of these systems, relying on deep learning techniques, require substantial amounts of training data. Notably, the majority of existing QA systems and datasets are predominantly available in English. The Stanford Question Answering Dataset (SQuAD) [22] is a well-known English QA dataset, consisting of approximately 100,000 questions based on Wikipedia articles. Other English QA datasets include MS Marco [23] and WikiQA [24], which were compiled by sampling questions searched on the Bing search engine, containing 100,000 and 3,000 samples, respectively. The Natural Questions dataset [25] comprises 300,000 examples gathered from questions searched on the Google search engine. Additionally, the NewsQA dataset, sourced from CNN news webpages, contains over 100,000 instances and focuses on training QA systems for English news domains.

While numerous QA systems have been developed for non-English language, the scarcity of large-scale datasets poses a challenge. Consequently, the initial step in building a QA system for a non-English language involves creating a dataset specifically tailored to that language. Two main approaches are commonly employed: (1) translating existing English QA datasets such as SQuAD using machine translation, and (2) constructing native QA datasets from scratch.

Some researchers have opted to translate SQuAD into different languages and utilize the translated dataset to build QA systems. For instance, Carrino et al.[26] translated SQuAD into Spanish and developed a Spanish QA system. Mozannar et al.[27] employed machine translation to convert 48,000 instances from SQuAD into Arabic, subsequently building an Arabic QA system. Lee et al.[28] performed machine translation of SQuAD into Korean and then developed a Korean QA system. Croce et al.[29] utilized semi-supervised translation techniques to create an Italian QA system by translating SQuAD into Italian.

Alternatively, some researchers have focused on building native QA datasets for target languages and employed them to develop QA systems. Efimov et al.citeefimov2020sberquad created a native QA dataset consisting of 50,000 instances for the Russian language, resulting in the development of a Russian

QA system. Shao et al.[30] developed a Chinese QA system using a native Chinese QA dataset comprising 30,000 instances. Similarly, Lim et al. [31] and Keraronb et al.[32] constructed QA systems for the Korean and French languages, respectively.

However, there are relatively few works dedicated to building QA systems for the Persian language. Veisi and Shandi[33] developed a Persian medical QA system to address questions related to diseases and drugs. Abadani et al[34] translated SQuAD into Persian and built a general Persian domain QA system. Kazemi et al[35] created PersianQuAD, a large-scale Persian QA dataset comprising approximately 20,000 questions, and implemented a QA system for evaluation purposes. Abadani et al[34] employed machine translation to convert a portion of the SQuAD dataset into Persian. They introduced a native answer selection dataset and a deep learning-based approach for Persian QA. Boreshban et al.[36] constructed a religious QA dataset for answering questions in the Persian religious domain, utilizing it to develop a QA system. Mozafari et al[37] and Kazemi et al[38] used deep learning methods to build Persian QA systems.

2.3 Evaluation of ChatGPT on QA Systems

In this section, we investigate the researches on evaluation of ChatGPT performance on question answering tasks. Omar et al.[39] compared traditional question answering systems (i.e. KGQAN and EDGQA) with ChatGPT for knowledge graphs. Several experiments were conducted on four real-world knowledge graphs over different applications. The authors discussed the advantages and disadvantages of each approach and proposed future directions for improving knowledge graph chatbots. The authors prompted ChatGPT in three modes: Default, Follow-up asking the entire list of response and Excel, requesting the answers in structured format. The QAS performance is evaluated in terms of determinism considering precision, recall and F1 score. The chatbot capabilities to produce correct and fluent answers is evaluated in terms of robustness, explainability and question understanding, through human manual assessment process. The paper suggests that ChatGPT provides more robust and explainable conversations compared to traditional QASs, but there are challenges to overcome, such as ensuring correctness, incorporating recent information and domain independency in responding.

In a most recent study, Tan et al.[40], developed an evaluation framework to analyze the reasoning capability of ChatGPT for answering complex questions. In the suggested framework, the performance of ChatGPT in question answering task is compared to several large language models like GPT-3 and GPT-3.5 upon 8 real world knowledge based datasets. Unlike our current study, the authors evaluated the capability of ChatGPT to answer the questions using its internal knowledge from two point of view, functionality and reliability. The former was assessed as the semantic parsing ability of ChatGPT on complex questions and the latter was assessed through compositional reasoning processes over its internal knowledge. This procedure facilitates automatic evaluation of QA related metrics such as exact match. As a result, the paper utilized both automatic and human evaluation for investigating the performance of ChatGPT, which can be considered as an advantage over completely manual evaluation methods. This study found that, the performance of ChatGPT is overall better than other model in monolingual QA except for multi-hob, star-shaped reasoning as well as number and time based answers. Furthermore, it was revealed that ChatGPT outperforms other models on multilingual tests.

In their recent study, Guo et al.[41] present the Human ChatGPT Comparison Corpus (HC3), a dataset consisting of question-answer pairs in various domains such as finance, medicine, law, psychology, and open domain questions. The authors aim to compare the answers generated by ChatGPT with those provided by humans, using both Turing and Helpfulness tests. For the Turing test, two groups of individuals, including experts and amateurs, were asked to differentiate between answers generated by ChatGPT and those given by humans. The evaluation metric used in this test was the proportion of correctly identified ChatGPT-generated answers. In the Helpfulness test, human experts were requested to assess the helpfulness of ChatGPT answers compared to human answers for each question. The evaluation measure for the Helpfulness test was the proportion of instances where ChatGPT answers were deemed more helpful than human answers. The study conducted both qualitative and quantitative analyses on the results obtained from ChatGPT. The authors observed implicit differences between the responses generated by humans and ChatGPT. However, no comparison was made with existing question-answering systems.

Again, Pichappan et al.[42] investigated ChatGPT as a question answering tool with a focus in text analysis of the answers. The authors conducted an experiments entering 27 random question including factual, interpretative, evaluative and rhetorical to ChatGPT and evaluating the answers through human and statistical evaluation procedures. For evaluating the answers, the output of ChatGPT are subjected to normal and AI similarity measures as well as various NLP parameters namely the number of wordiness instances, rephrasing and rearranging, text inconsistency occurrence, engagement, unclear sentences and determiners.

These parameters are taken into account for assessing the effectiveness of the ChatGPT from text analysis point of view. They concluded the poor originality for ChatGPT and claimed that ChatGPT performs like an intelligent plagiarist by not providing the source of original texts used to generate answers. Furthermore, authors found that the precision of ChatGPT on linguistic parameters still need significant improvements and it also fails to meet users' requirements, completely.

To sum up, in enumerated researches, there have been no examinations of ChatGPT's ability to extract answers from provided paragraphs. When utilized as a question-answering system (QAS) without any accompanying text for answer extraction, ChatGPT relies solely on its acquired knowledge to generate responses, which may be influenced by training biases. Moreover, the issue of hallucination arises when ChatGPT struggles to accurately identify the answer to a specific question. It seems that there is still an open issue on evaluating the ChatGPT performance in answer extraction from given paragraphs as a key challenge of QASs.

Table 1 summarizes the evaluation of ChatGPT in various studies along with the key differences in our work:

3 The Evaluation Framework

To assess the accuracy and efficiency of the responses obtained from the ChatGPT system, an evaluation framework has been employed (see Figure 1). This framework comprises three main components: Prompt builder, ChatGPT, and Answer evaluator. The overall structure of a QA dataset is in the form of a set of triples (P, Q, A), in which P refers to a textual paragraph, Q denotes a question about the paragraph, and A stands for the corresponding answer. In some datasets, the answer to the question may not exist in the paragraph, and an additional field indicates that the answer does not exist in the corresponding paragraph. We refer to these two dataset types as AQA: Answerable Question Answering and AUQA: Answerable/Unanswerable Question Answering, respectively. For each triple of the dataset, P_i and Q_i are entered into the Prompt builder section, and a prompt is prepared and delivered to ChatGPT to ask Q given P. Then, the generated answer of ChatGPT along with the gold answer A are passed to the Answer evaluator component to analyze the performance using predefined evaluation metrics.

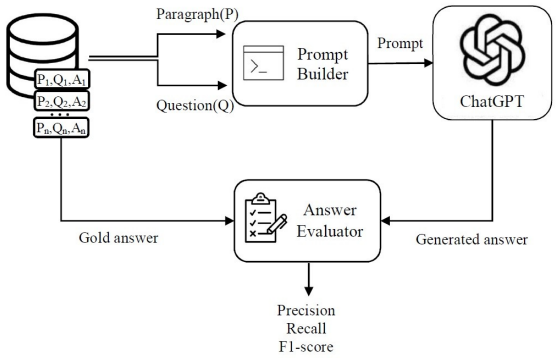

Figure 1. The Architecture of the Evaluation Framework.

3.1 Prompt Builder

The prompt builder is responsible for generating a proper prompt regarding the available data for feeding to ChatGPT. The prompt used for evaluating the questions was obtained through trial and error, aiming to achieve the highest accuracy from the GPT-3.5 model. This proposed prompt attempts to address all limitations, including the number of allowable characters returned by the GPT-3.5 model and the model's sensitivity to key words. Changing the prompt may result in longer or incorrect responses, directly impacting evaluation metrics such as precision and recall, and consequently affecting the F1 score. Therefore, two main prompts are designed for the answerable question answering datasets, in which all the questions have their answers present in the given paragraphs and answerable/ unanswerable question answering datasets which can also contain questions whose answers are not present in the given texts.

- **AQA: Answerable question answering:** During the questioning process, an attempt has been made to only ask questions for which the answers exist in the text. The prompt designed for AQA is as follows:
  `"what is the answer of this question '" + Q + "' based on this paragraph '" + P +"'? say only the answer without saying the full sentence"`
- **AUQA: answerable/ unanswerable question answering:** In the AUQA dataset, due to the presence of questions that do not have corresponding answers in the text, a two-step process is employed to obtain answers for the questions. In the first step, the text and question from the dataset are fed to the language model. the model is queried to determine whether the question exists in the provided text or not. The prompt for this step is as follows: `"Given the question '" + Q + "', determine if the following paragraph contains the corresponding answer: " + P +`

Table 1. Comparison of ChatGPT Evaluation Studies and Our Work.

Study	Focus Area	Evaluation Aspects/Methods	Differences in Our Work
[1], [2],[3]	General evaluation of ChatGPT	Various abilities (e.g., reasoning, understanding)	Our work specifically focuses on QA tasks
[4], [5]	Mathematical problem solving	Problem solving tasks	Our work addresses broader QA tasks
[6]	Understanding abilities	Comprehension tasks	Our study involves answering questions based on text
[7], [8], [9]	Code bug fixing	Coding tasks	Our research does not involve coding
[11], [12], [13]	Translation tasks	Translation accuracy	Our focus is on QA, not translation
[16]	Out-of-distribution behavior	Response analysis	We focus on QA within provided texts
[17]	Open information extraction	Extraction accuracy	We evaluate ChatGPT's QA from specific paragraphs
[39]	Knowledge graph QA comparison	Question answering ability in terms of Precision, Recall and F1 score	Our work evaluates QA from textual paragraphs, not knowledge graphs
[40]	capability to answer using the internal knowledge	functionality and reliability	We focus on paragraph-based QA extraction while this study evaluate the internal knowledge of ChatGPT
[42]	answer's text analysis	Normal and AI similarity measures, the number of wordiness instances, rephrasing and rearranging, text inconsistency, engagement, unclear sentences and determiners	the study analyses the NLP parameters of the answer
[41]	Human vs. ChatGPT QA comparison	Turing test, helpfulness test	Our study includes text-based QA evaluation not covered by Guo et al.
Current Research	**QA tasks from textual paragraphs**	**Accuracy, relevance, comprehensiveness, human evaluation**	**evaluating ChatGPT's QA performance using specific textual paragraphs, employing both quantitative metrics and human evaluation to assess accuracy, relevance, and comprehensiveness in detail.**

`" say only 'Yes' or 'NO'"`

If the answer to the question is "Yes,", then we need to ask again to find the answer in the paragraph. The prompt for this second step is as follows: `"what is the answer of this question'" + Q + "'based on this paragraph'" + P +"'? say only the answer without saying the full sentence"` It is important to note that only questions that are present in the text are used for evaluation. If a question does not exist in the text, nothing is asked from the model for that particular question during evaluation. As we see in Table 2, NewsQA includes questions related to hallucinations. We give questions that we are sure have no answer to ChatGPT, and if answers are returned for those questions, it indicates a hallucination. If the text contains an answer to the asked question, it should return the answer; otherwise, it should return 'None.' In contrast, all SQuAD 1.1 questions can be found within the paragraphs provided. However, SQuAD 2.0 may include questions that are not present within the paragraphs, making it distinct in terms of question content.

As depicted in Table 2, the prompt builder encompasses two distinct types. This tailored approach ensures a more nuanced understanding of the data and contributes to the precision of generated responses, resulting in a better exact match.

Table 2. Example of Prompt Builder.

Dataset	Paragraph	Dataset Type	Prompt	GPT Response
NewsQA	NEW DELHI, India (CNN) – A high court in northern India on Friday acquitted a w...	AQA	what is the answer of this question "What is the number of children murdered based on this paragraph "A high court ... ? say only the answer without saying the full sentence	N/A (No specific answer provided)
SQuAD 2.0	The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Norma...	AUQA	what is the answer of this question "In what country is Normandy located?" based on this paragraph "The Normans..."? say only the answer without saying the full sentence	No
SQuAD 1.1	Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defe ...	AQA	Which NFL team represented the NFC at Super Bowl 50?	Denver Broncos

3.2 ChatGPT

ChatGPT is one of OpenAI's important products with the capability of human-like language interaction. This product utilizes the GPT-3.5-turbo language model to provide responses to posed questions. In this research, an API connected to this language model has also been employed. Due to limitations on the number of requests and the number of tokens sent, significant challenges arose in obtaining information from the model. To overcome these challenges, various methods were used.

One of the methods used in this research is prompt engineering. In this study, we aimed to extract responses from ChatGPT in such a way that the evaluation metrics could be correctly applied when querying the datasets. For example, for any question that ChatGPT could not answer correctly, we forced it to return the word NONE instead of an incorrect or similar answer. For instance: "What is the answer to this question: question_text in this text: story_text? If there is no answer, just return NONE."

Note that the OpenAI API has capabilities such as setting the temperature parameter (to determine the randomness of the response) and defining roles (this parameter specifies in what role the model should respond to your question, such as a specialist, an ordinary person, a professor, a student, etc.). By default, the role was set to assistant. Additionally, the temperature parameter was set to its default value to minimize the generation of creative responses as much as possible.

Furthermore, the questions were asked in separate API calls to ensure that ChatGPT's responses did not influence subsequent answers. Additionally, due to the limitations we faced, we were forced to query the model in 3-second intervals (as with the free plan, only 20 requests per minute can be sent). Moreover, given the limitations of Colab for using hardware resources, we saved the answers on Google Drive at each query stage to ensure that extracted responses were not lost in case the runtime was reset. For example, to address the issue of token limitations, prompt engineering and character limit considerations were used. To tackle the time constraint challenge, we resolved it by adjusting the request rate to 3 requests per minute (one request every 20 seconds).

3.3 Answer Evaluator

In order to evaluate the responses generated by ChatGPT, we initiate the process by normalizing both the ground truth sentence, which encapsulates the correct answer to the question within the dataset, and the response from ChatGPT. This normalization entails the removal of punctuation marks, special characters, and white spaces, along with converting all characters to lowercase. Following this preprocessing step, we proceed to compare these two sentences at the word level. We employ metrics such as Exact Match, F1 score, and Recall to quantitatively assess the level of correspondence between the generated response and the ground truth. These measures serve as valuable benchmarks for evaluating ChatGPT in comparison to QASs that have operated on similar datasets. This comprehensive evaluation methodology enables us to make informed comparisons and draw meaningful conclusions about the performance of ChatGPT in relation to its predecessors.

The equations for these metrics are as follows:

$$\text{Precision} = \frac{|\text{Correct Words In Response}|}{|\text{Total Words In Response}|} \quad (1)$$

$$\text{Recall} = \frac{|\text{Correct Words In Response}|}{|\text{Total Words In Ground Truth}|} \quad (2)$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Exact Match} = \frac{|\text{PredictedAnswers} \cap \text{CorrectAnswers}|}{|\text{TotalQuestions}|} \times 100 \tag{4}$$

By using these metrics, we can accurately measure the performance of ChatGPT in handling QA tasks, providing a clear comparison with similar QA systems.

4 Experimental Results

In this section, we first introduce the datasets used for evaluation. Next, we introduce the evaluation metrics employed. Subsequently, we present the obtained results and compare them with the results of other models. Finally, the impact of prompt variations on the conclusions is examined. For this experiment, we employed Google Colab for the collection and analysis of data.

4.1 Datasets

We aim to evaluate ChatGPT's performance in question-answering (QA) tasks, with a particular emphasis on its effectiveness in handling queries across diverse languages, specific domains, and datasets, including those with unanswerable questions.

Numerous QA datasets are available for assessing question-answering systems in the English language. In our investigation, we chose to use three renowned datasets to assess ChatGPT's proficiency in English QA tasks. These datasets include the Stanford Question Answering Dataset (SQuAD), encompassing both its initial version SQuAD 1.1 [22] and SQuAD 2.0 [43], as well as NewsQA [44]. For assessments related to the Persian language, we utilized the PersianQuAD dataset [35]. Each dataset presents unique characteristics and challenges, enabling a comprehensive evaluation of the model's capabilities. We also describe which dataset is used for each type of evaluation.

- **SQuAD 1.1:** SQuAD 1.1 is widely recognized as one of the preeminent QA datasets in the English language. It boasts an extensive collection of over 100,000 questions, thoughtfully curated with contributions from crowdworkers who generated questions based on specific Wikipedia articles. Annotators were tasked with formulating queries based on provided article paragraphs and specifying corresponding answers. This dataset is mostly used for the Answer-Definite questions.
- **SQuAD 2.0:** SQuAD 2.0 introduces an additional layer of complexity by incorporating over 50,000 deliberately unanswerable questions, intricately designed by crowdworkers to closely resemble answerable counterparts. This dataset is used for Answer-Possible questions, where most answers are absent in the paragraph. It is suitable for evaluating the quality and efficiency of answering questions where the answers are present in the text.
- **NewsQA:** NewsQA is meticulously crafted through a three-stage process involving different roles: questioners, answerers, and validators. Initially, questioners formulate questions based solely on article highlights and headlines. Answerers then assess these crowdsourced questions alongside the full article to derive accurate responses. In the final stage, validators meticulously evaluate the article, question, and proposed answers to select the most appropriate response or reject inadequate ones. This rigorous evaluation focuses exclusively on questions with verifiable answers within the text.

  The exceptional quality and precision of the dataset render it indispensable for evaluating language models, particularly in identifying instances of hallucination, where models generate answers lacking textual grounding or accuracy. NewsQA's design ensures robust validation in detecting such errors, vital for enhancing natural language understanding models. By training on this dataset, models can effectively distinguish between well-supported answers and those that are speculative or incorrect, thereby advancing their capability to deliver reliable responses rooted in textual evidence. NewsQA thus stands as a pivotal resource for refining and assessing the performance of language models in real-world question-answering scenarios while minimizing erroneous output.
- **PersianQuAD:** PersianQuAD represents a pioneering initiative and marks the inception of a comprehensive, large-scale native machine reading comprehension dataset designed specifically for Persian language question-answering. This dataset consists of approximately 20,000 meticulously curated questions crafted by native annotators.

  To construct PersianQuAD, annotators were tasked with generating questions based on paragraphs extracted from Persian Wikipedia articles. Each question was designed to have its answer explicitly stated within the paragraph text. This meticulous process ensures that PersianQuAD provides a robust framework for evaluating the performance of question-answering models tailored to the Persian language.

  Given its carefully annotated nature and fo-

cus on Persian language comprehension, PersianQuAD serves as an ideal dataset for assessing the efficacy and accuracy of question-answering models operating in Persian. It enables researchers and developers to train and evaluate models that can effectively comprehend and respond to questions posed in Persian.

The thoughtful selection of these datasets, along with our rigorous evaluation criteria, forms a robust foundation for the comprehensive assessment of ChatGPT's performance across diverse linguistic contexts within the realm of QA tasks. This endeavour aims to provide valuable insights into ChatGPT's capabilities and limitations, contributing to a deeper understanding of its effectiveness.

4.2 Metrics

The two common evaluation metrics used for assessing QA systems are: "Exact Match" and "F_1 score"[22]. The same metrics are employed in our research. In this study, we also use the Recall measure.

- Exact Match: This metric calculates the percentage of predicted answers that precisely match any of the correct candidate answers.
- (Macro-averaged) F_1 score: This metric determines the average overlap between the predicted answers and the correct candidate answers. To calculate the overlap, both the predicted and candidate answers are treated as bags of words, disregarding word order. For each question, the F_1 score is determined as the maximum F_1 score over all of its candidate answers. The "(Macro-averaged) F_1 score is the mean of the F_1 scores across all questions.
- Recall: Recall measures the QA system's ability to find all the relevant answers from a given set of documents or data sources. It is calculated as the ratio of the number of correctly retrieved relevant answers to the total number of relevant answers.

By utilizing these three evaluation criteria for each of the four datasets used in the evaluation, a comprehensive assessment of the Chat GPT system's performance is achieved. These metrics enable a quantitative evaluation of the model's accuracy, completeness, and overall effectiveness in answering questions. All the results presented in the tables are expressed in percentages.

4.3 Comparing ChatGPT With Language Models on the Question Answering Task

In this section, a comprehensive examination of various datasets has been undertaken, employing the GPT-3.5-Turbo language model as a question-answering system. The resulting insights stem from a meticulous comparison between the GPT-3.5-Turbo language model and other well-established language models. These models have undergone extensive testing and evaluation across diverse datasets, including SQuAD, NewsQA, and PersianQuAD, serving as robust benchmarks for gauging the capabilities of the GPT-3.5-Turbo language model within a broader context.

To ensure a nuanced evaluation of the GPT-3.5-Turbo, we strategically selected popular question-answering models tailored for each dataset. For SQuAD 1.1, our chosen models include LUKE, XLNet, and SpanBERT.

In Table 3, the values of Exact Match and recall for these preceding question-answering models surpass those of GPT-3.5-Turbo. This discrepancy can be attributed to the specialized development of these models explicitly for question-answering tasks.

In Table 4, which focuses on SQuAD 2.0, we opted for Retro-Reader, Retro-Reader on ALBERT, and ALBERT. Notably, the values of Exact Match and recall for GPT-3.5 are lower than those of the other models, and they also fall short of the metrics observed in SQuAD 1.1. The rationale behind this disparity lies in the presence of unanswerable questions within SQuAD 2.0. In instances where the answer to a question is absent in the provided paragraph, GPT-3.5 generates responses, resulting in diminished Exact Match and recall values compared to SQuAD 1.1.

In Table 5, as observed, the dataset NewsQA comprises relatively lengthy stories, making the selection of relevant words challenging for each language model. This challenge is notably evident in the relevant terms and it may contribute significantly to the model's lower recall. The reduction in relevant words will inevitably result in shorter final model responses compared to the answers present in the dataset. Additionally, the inference of certain questions from various parts of the text could be another factor contributing to the lower exact match (EM) scores. Analyzing the results from the table, it becomes apparent that achieving high performance on NewsQA is a complex task, requiring models to navigate through lengthy narratives and effectively capture relevant information for accurate responses. This insight sheds light on potential avenues for improving the performance of language models in handling comprehension tasks on datasets with extended and context-rich content.

In Table 6, we shift our focus to evaluating the capabilities of GPT-3.5-Turbo in question-answering tasks for languages other than English. To accomplish this, we conduct a comparative analysis, pitting GPT-

Table 3. Assessing the Effectiveness of Various Language Models on SQuAD 1.1.

Model Name	Exact Match	Recall
LUKE [45]	90.2	95.4
XLNet [46]	89.898	95.080
SpanBERT [47]	88.8	94.6
ChatGPT(GPT-3.5-Turbo)	44.4	87

Table 4. Assessing the Effectiveness of Various Language Models on SQuAD 2.0.

Model Name	Exact Match	Recall
Retro-Reader [48]	90.578	92.978
Retro-Reader on ALBERT [48]	90.115	92.580
ALBERT [49]	89.731	92.215
ChatGPT(GPT-3.5-Turbo)	18.121	65

Table 5. Assessing the Effectiveness of Various Language Models on NewsQA.

Model Name	Exact Match	Recall
RAG-original [50]	4.33	7.92
BERT+ASGen [51]	54.7	64.5
AMANDA [52]	48.4	63.7
DecaProp [53](Yi Tay et al. 2019)	53.1	66.3
ChatGPT(GPT-3.5-Turbo) for Questions with all validator-confirmed answers	41.07	46.70

3.5-Turbo against models specifically designed for the complexities of Persian Quad. The findings reveal a significant performance gap, with GPT-3.5 lagging behind its counterparts. This disparity stems from the fact that the Persian Quad model is purpose-built, having undergone rigorous training on the Persian Quad dataset and, importantly, on the Persian language itself. The Persian Quad model's concentrated focus on both dataset and language nuances clearly grants it a competitive advantage over GPT-3.5-Turbo's more generalized language understanding in the domain of non-English languages.

4.3.1 Assessing Performance Across Different Input Conditions and Languages

The SQuAD datasets, particularly SQuAD v1.0, hold a prominent position in the realm of question answering, serving as a benchmark for numerous researchers and their model development efforts. To evaluate ChatGPT's performance (GPT-3.5-Turbo) on these two datasets, two scenarios were considered: one where the paragraph is provided to the language model and the other where it is not. When the paragraph is not given, only the language model is asked to answer the corresponding question in the dataset. When the paragraph is provided in the prompt, there is a notable improvement in F1 score, Recall, and Precision. This clearly demonstrates ChatGPT's enhanced ability to identify accurate answers when given access to the relevant context. Table 7 shows the results.

Table 6. Results on PersianQuad.

Model Name	Exact Match	Recall
PersianQA [35]	78.8	82.97
ChatGPT(GPT-3.5-Turbo)	41	55

To assess ChatGPT's performance in non-English languages, the PersianQuAD dataset was selected. Similar to SQuAD 1.1, PersianQuAD ensures that answers to all questions reside within the corresponding paragraphs. The evaluation results are presented in Table 8, revealing a consistent decline in F1 score, Exact Match, and Precision compared to the SQuAD datasets. This observation suggests that ChatGPT's performance on non-English languages is not as good as its performance on English language. To evaluate ChatGPT's ability to handle hallucinations, the NewsQA dataset was employed.

Table 9 presents the results of ChatGPT on the NewsQA dataset, revealing its performance in terms of F1 score, Exact Match, and Precision.

Questions for the NewsQA and PersianQuAD datasets were created based on specific paragraphs. The evaluation metrics used are designed to measure the model's performance within these specific datasets and domains.

4.3.2 Impact of Prompt Modification

The quality of questions significantly affects the answers obtained from the language models. Using clear, simple, and concise prompts is crucial to improve response quality.

At the beginning, the questions were fairly complex. However, through ongoing refinement and simplification of the prompts, substantial improvements were achieved to obtain more accurate responses from the language model. This iterative process involved revising the prompts multiple times to better match the model's understanding and capabilities. Consequently, the questions became clearer and more straightfor-

Table 7. Comparison of Evaluation Metrics for SQuAD 1.1 and SQuAD 2.0.

Dataset	Scenario	F1 score (%)	Recall (%)	Precision (%)
SQuAD 1.1	Without Paragraph	1.92	1.98	2.05
	With Paragraph	67	87	62
SQuAD 2.0	Without Paragraph	2.51	2.53	2.72
	With Paragraph	56	65	54

Table 8. Evaluation Metrics for PersianQuad Dataset.

Metric	F1 score (%)	Recall (%)	Precision (%)
PersianQuAD	55	65	53

Table 9. Evaluation Metrics for All Questions of NewsQA.

Metric	F1 score (%)	Recall (%)	Precision (%)
All Questions	10.292	15.66	51.14

Table 10. Evaluation of Question Answering on SQuAD 2.0 for Different Prompts.

Method	Precision (%)	F1 Score (%)	Recall (%)
Step 1 + Step 2	27	37	88
Step 1	62	67	87

ward, resulting in improved accuracy in the responses generated by the model.

For SQuAD 2.0, some questions do not have answers. To improve this, a two-step evaluation process was implemented. One part of our research aimed to evaluate the accuracy with which ChatGPT answers the questions we pose. A type of question we asked the model is "If the answer to the question I gave you is present in the text, return 'Yes,' otherwise return 'No.'" Our reason for this approach was to specifically evaluate the questions in which we are sure the answer is in the text, without testing the model's internal knowledge, which might low evaluation metrics.

Step 1: `"Given the question " + question + ", determine if the following paragraph contains the corresponding answer: " + paragraph + " say only 'Yes' or 'No'"`

and if the returned response was "Yes", then we have:

Step 2: `"What is the answer to this question " + question + "based on this paragraph " + paragraph + "? Only provide the answer without full sentence."`

This two-step refinement significantly improved the model's performance, particularly for unanswerable questions. Importantly, the goal of this evaluation is not to assess the internal knowledge of ChatGPT but rather its ability to extract answers from a given passage. In cases where no answer is present in the text, the model is expected to correctly respond with "no answer."

As demonstrated in Table 10, changing the prompt to more relevant and direct words, improved the model's ability to extract accurate answers with higher F1 score.

4.3.3 Evaluation on Question Types (how-, what-, who-, when-, where- and list questions)

ChatGPT performance (GPT-3.5-turbo) in various types of questions in the SQuAD 1.1 and SQuAD 2.0 datasets was evaluated by categorizing initial question words such as 'what,' how,' when,' where,' who,' which and 'why'. Each category was assessed using F1 score, precision and recall metrics to gauge the model's ability to provide accurate answers.

The results, depicted in Figures 2, 3, and 4, indicate that ChatGPT excels in answering factual questions (what, when, where, who, which), demonstrating higher scores in these categories across all metrics. However, it struggled with questions starting with 'why' and 'how,' which require deeper reasoning and understanding of causal relationships. This disparity in performance underscores the complexity of these types of inquiry for current language models, suggesting a need for more sophisticated evaluation methods tailored to such complex questions.

Figure 2 illustrates the precision comparison. Here again, ChatGPT shows higher precision scores for factual questions compared to 'why' and 'how' questions. Precision is particularly high for 'who' and 'which' questions in SQuAD 1.1, but drops in SQuAD 2.0. Figure 3 presents the recall comparison. The trends are consistent with the F1 score and precision results, with better recall for factual questions and lower recall for the "why" and "how" questions.

Figure 4 shows the F1 score comparison, highlight-

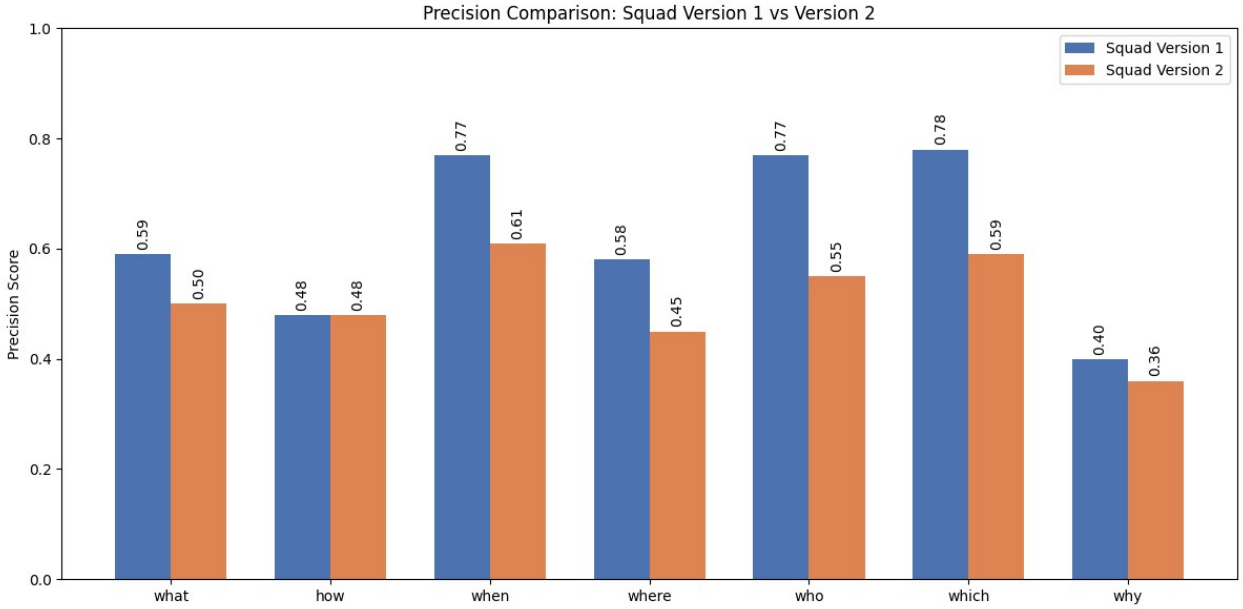

Figure 2. Histogram of Computed Precision Across Various Question Types Within the SQuAD 1.1 and SQuAD 2.0 Dataset.

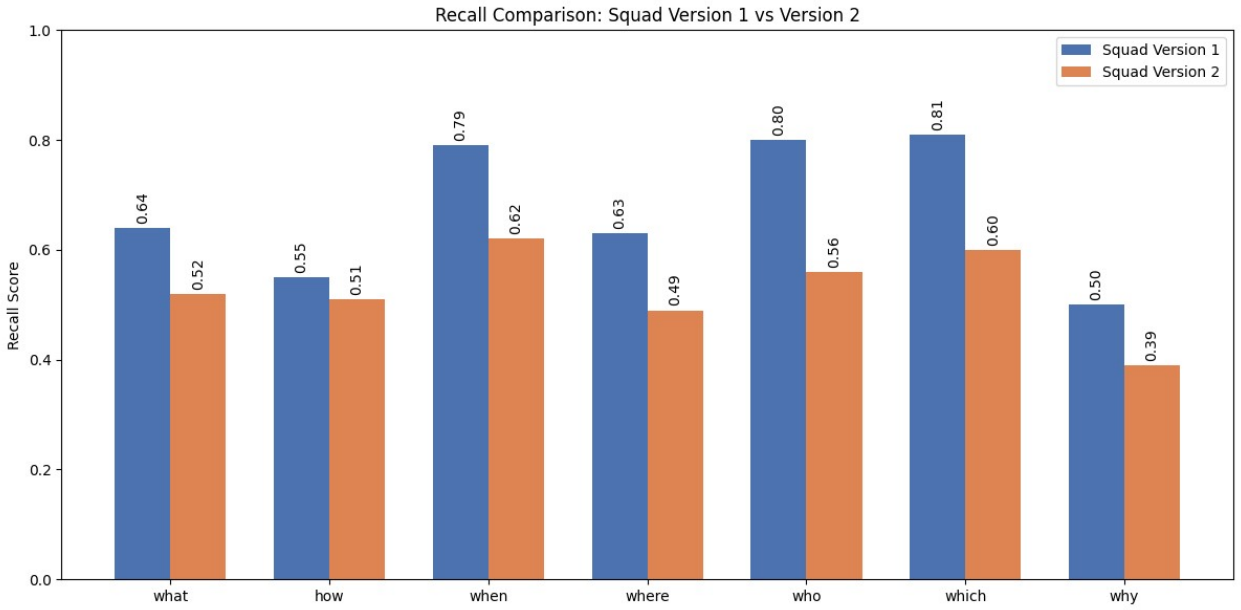

Figure 3. Histogram of Computed Recall Across Various Question Types Within the SQuAD 1.1 and SQuAD 2.0 Dataset.

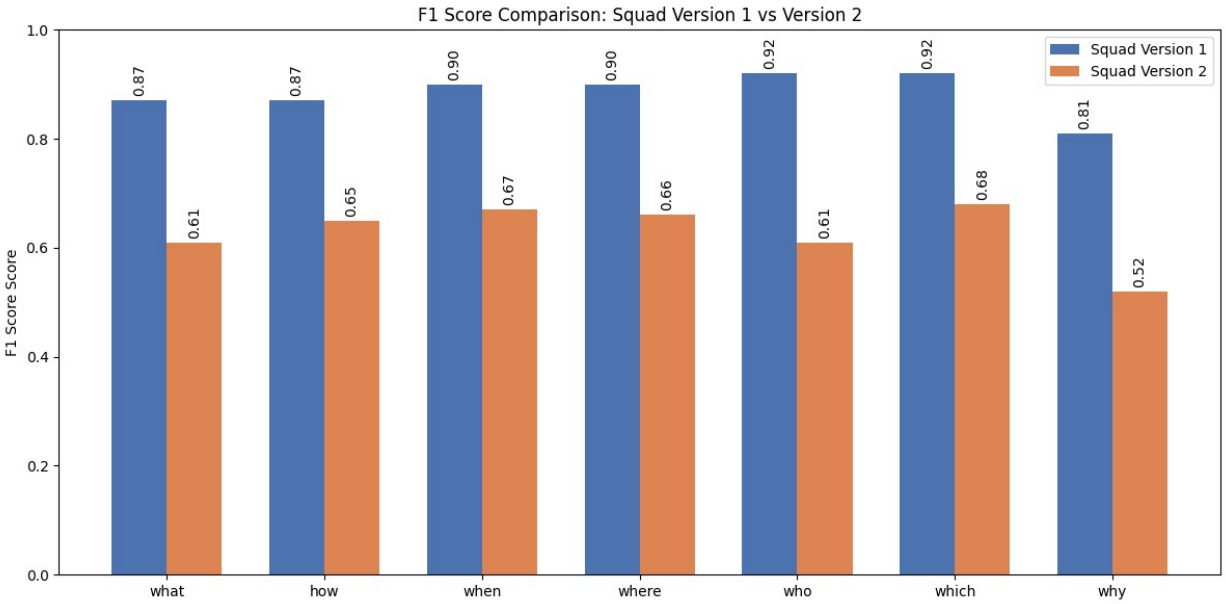

Figure 4. Histogram of Computed F1 Score Across Various Question Types Within the SQuAD 1.1 and SQuAD 2.0 Dataset.

ing that ChatGPT performs significantly better on 'which,', 'when,' 'where,' and 'who' questions in both SQuAD 1.1 and SQuAD 2.0 datasets. The F1 scores for the "why" and "how" questions are notably lower.

These visual representations highlight ChatGPT's strengths in factual question answering and its challenges with reasoning questions. To enhance evaluation robustness, future analyzes could include comparisons with baseline models, explore alternative evaluation methodologies such as human assessments or advanced semantic measures for the 'why' and 'how' questions, and conduct error analyses to pinpoint specific areas for improvement. Such approaches would provide deeper insights into ChatGPT's capabilities and limitations across diverse question types, guiding advancements in natural language understanding and reasoning tasks.

In conclusion, while ChatGPT demonstrates competence in factual question answering, its performance varies significantly between different question types. Addressing the complexities associated with reasoning questions like "why" and "how" remains a crucial area for improvement. By refining evaluation strategies and leveraging insights from comparative analyses, researchers can advance the development of language models capable of more nuanced understanding and precise responses in a wide range of inquiries.

4.3.4 Evaluation on ChatGPT's Hallucination

It is crucial to address the presence of questions in datasets that are not explicitly mentioned in the provided text. To mitigate this challenge, an evaluation framework ensures that ChatGPT refrains from generating responses when questions are absent from the text. This measure aims to prevent incorrect responses that could bias the evaluation metrics.

Within this framework, we investigate the phenomenon known as "hallucination," where ChatGPT generates responses that appear plausible but lack a factual basis. The process involves querying the model about the existence of a given question in the provided text. If the model responds to a question confirmed as absent, the generated response is categorized as a hallucination.

In our implementation, two key fields, namely isAnswerAbsent and isQuestionBad, are used to assess the presence of an answer in the text. A value of 1 in both fields confirms that the question is definitively absent from the text. The prompt utilized for this evaluation is:

```
what is the answer to this question: {question_text}
in this text: {story_text} if there is no answer,
just return NONE
```

Further investigation based on the results of testing 5,598 questions from the NewsQA dataset (approximately 1% of the dataset) indicates that 91.31% of these questions had at least one validator, suggesting potential instances of incorrect answers. Despite

the absence of a clear answer, the model attempted responses based on contextual clues provided by validators. Conversely, 8.68% of the questions lacked any validators, indicating responses that may be considered hallucinations.

This evaluation process aims to meticulously identify and analyze instances of hallucination, ensuring that generated responses are grounded in the factual context of the given text. Such rigorous scrutiny significantly enhances the reliability and trustworthiness of ChatGPT's performance, particularly in natural language understanding tasks.

Table 11. Evaluation of Question Answering on SQuAD 2.0 for Different Prompts.

Method	Precision (%)	F1 Score (%)	Recall (%)
Step 1 + Step 2	27	37	88
Step 1	62	67	87

In our implementation, questions are categorized based on two key fields: 'is answer absent' and 'is question bad'. If either 'is answer absent' or 'is question bad' is true, the question is marked as Type X. This indicates that the model attempted to answer a question that was absent or poorly formulated, which can lead to hallucination. Conversely, if both 'is answer absent' and 'is question bad' are false, the question is marked as Type J, indicating that the model is expected to provide a valid response based on the text.

To illustrate the evaluation process and its outcomes, we present a few examples of generated responses in Table 12.

4.3.5 Evaluation on Question Complexity

To evaluate ChatGPT's ability to answer questions of varying difficulty, we focused on calculating the F1 score across multiple datasets: NewsQA, PersianQuad, SQuAD 1.1, and SQuAD 2.0. The core challenge lies in measuring how well ChatGPT matches questions with their correct answers within given paragraphs. This is crucial because the closer a question is to its answer in terms of wording and context, the easier it is for the system to respond accurately. To quantify this closeness, we used the Jaccard Coefficient, which assesses similarity based on shared words between a question and its answer. Kazemi et al.[35] also used Jaccard for this study.

Formally, the Jaccard Coefficient, denoted by J(Q, A), is calculated as follows:

Table 12. Sample Model-Generated Responses.

Story Text	Question	Answer	gpt_answer	Type
(CNN) – Pope John Paul II used to beat himself with a belt and sleep on a bare	what Journalist compares flagellation to fasting or other?	NONE	The journalist mentioned in the text is David Gibson.	x
(CNN) – Pope John Paul II used to beat himself with a belt and sleep on a bare	what did he do	NONE	The answer to the question "What did he do?" in this text is: Pope John Paul II used to beat himself with a belt and sleep on a bare floor to bring himself closer to Christ.	x
(CNN) – England international football baller Steven Gerrard was found not guilty o	What did Gerrard admit doing?	punching businessman Marcus McGhee.	The answer to the question "What did Gerrard admit doing?" is throwing three punches.	j
(CNN) – England international football baller Steven Gerrard was found not guilty o	Who did Gerrard admit to punching?	businessman Marcus McGhee.	Gerrard admitted to punching businessman Marcus McGhee.	j

$$J(Q, A) = \frac{|Q \cap A|}{|Q \cup A|} \tag{5}$$

As outlined in the Jaccard Coefficient calculates similarity by considering the ratio of common words shared between the question and answer sentences to the total number of words in both the question and answer. With the Jacard Coefficient ranging between 0 and 1, a value of 0 signifies no lexical similarity between the question and the sentence, while a value of 1 indicates that the sentence encompasses all the words present in the question.

This metric ranges from 0 to 1, where 0 indicates no overlap between question and answer, and 1 indicates complete overlap.

We analyzed ChatGPT's performance by plotting F1 scores across different ranges of Jaccard similarity values. Figures 5-8 illustrate these scores across various datasets, revealing intriguing patterns. For instance, in SQuAD 1.1 (Figure 5), ChatGPT achieved its highest F1 score of 0.83 when handling questions with a Jaccard similarity of 0.5, showcasing strong performance on moderately complex questions.

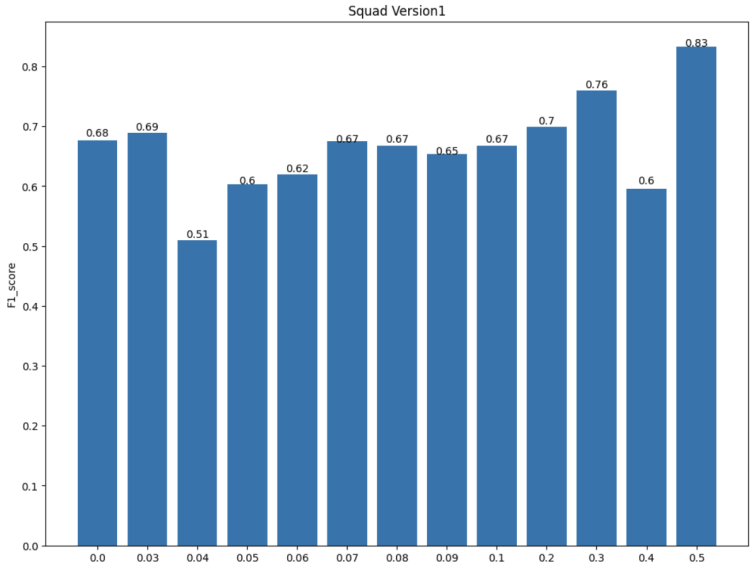

Figure 5. Impact of Question Similarity on F1 Score Performance in SQuAD 1.1.

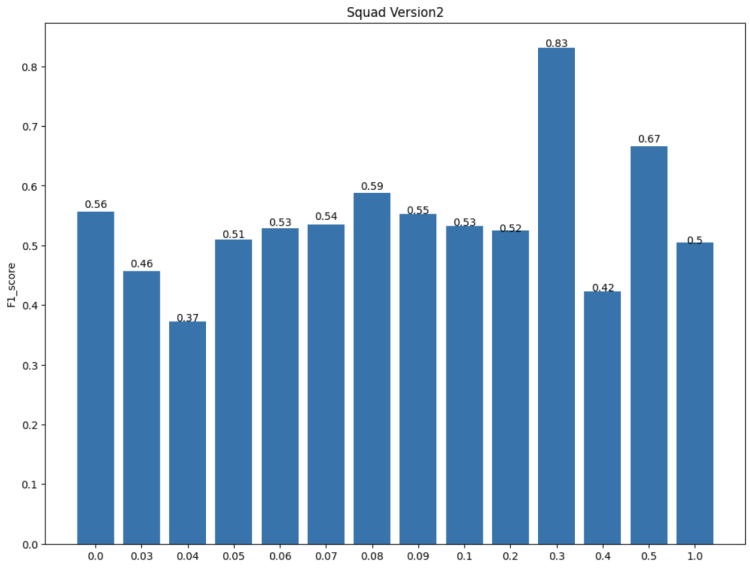

Figure 6. Impact of Question Similarity on F1 Score Performance in SQuAD 2.0 Dataset.

In SQuAD 2.0 (Figure 6), ChatGPT excelled with an F1 score peak of 0.78 on questions with a Jaccard similarity of 0.3, suggesting robust performance on questions with lower lexical overlap.

In the PersianQuad dataset (Figure 7), ChatGPT reached a top F1 score of 0.78 on questions where the Jaccard similarity was 0.09, highlighting its capability in dealing with questions requiring more nuanced understanding.

Figure 8 depicts findings from the NewsQA dataset, where we transformed the difficulty values into twenty-point intervals for clarity. Here, we observed a cyclic pattern in F1 scores as question difficulty varied. This cyclic behavior implies that ChatGPT's performance fluctuates with different levels of question complexity, rather than a linear increase or decrease.

These insights deepen our understanding of ChatGPT's performance across diverse question types and difficulty levels, underscoring its strengths and areas for improvement in natural language understanding tasks.

5 Conclusions

In this study, we conducted an analysis of ChatGPT as a Question Answering System (QAS) and com-

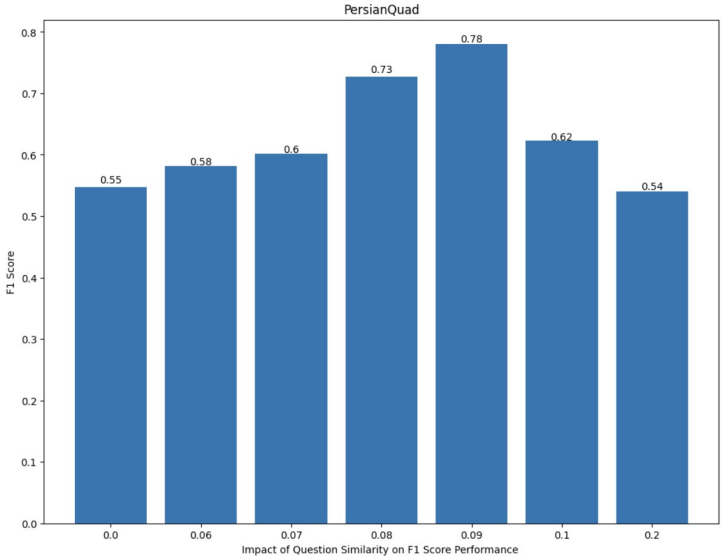

Figure 7. Impact of Question Similarity on F1 Score Performance in PersianQuad.

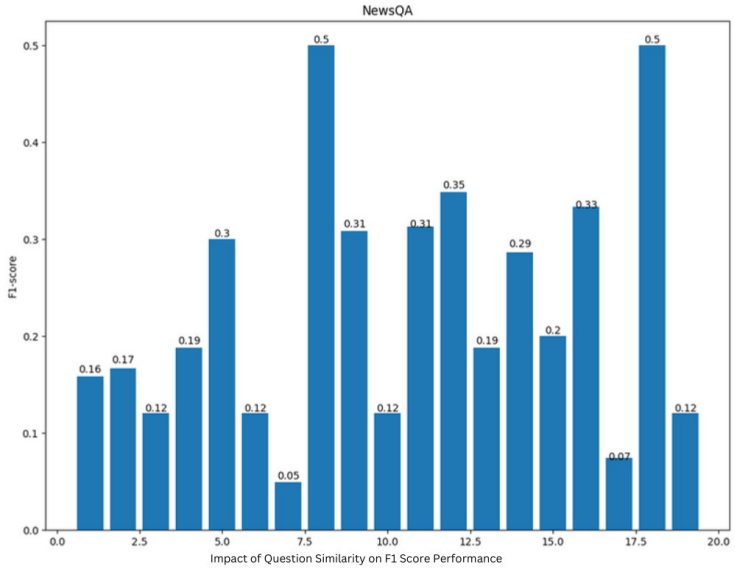

Figure 8. Impact of Question Similarity on F1 Score Performance in NewsQA.

pared its performance with existing QA models. Our evaluation focused on the model's ability to extract answers from provided paragraphs, as well as its performance in scenarios without surrounding passages. We explored various aspects, including response hallucination, question complexity, and the impact of context on performance.In conclusion, our comprehensive evaluation of ChatGPT as a QAS has provided valuable insights into its strengths, limitations, and areas for improvement. Leveraging well-known Question Answering (QA) datasets in both English and Persian languages, we employed metrics such as F-score, exact match, and recall to assess ChatGPT's performance.

Our findings indicate that ChatGPT, while demonstrating competence as a generative model, faces challenges in question answering compared to task-specific models. Context proves to be a crucial factor, with the model exhibiting improved performance when provided with surrounding paragraphs for answer extraction. Prompt engineering, particularly in the form of two-step queries, enhances precision, especially for questions lacking explicit answers in provided passages. We observed ChatGPT's proficiency in answering simpler, factual questions, highlighting its strengths in handling straightforward queries. However, challenges arise with more complex "how" and "why" question types. The evaluation also unveiled instances of hallucinations, where ChatGPT provided responses to

questions without available answers in the provided context.

References

[1] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.ijcnlp-main.45. URL https://aclanthology.org/2023.ijcnlp-main.45/.

[2] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023. doi:10.48550/arXiv.2302.06476. URL https://arxiv.org/abs/2302.06476.

[3] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*, February 2023. doi:10.48550/arXiv.2302.10724. URL https://arxiv.org/abs/2302.10724.

[4] Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023. doi:10.48550/arXiv.2301.13867. URL https://arxiv.org/abs/2301.13867. NeurIPS 2023 Datasets and Benchmarks.

[5] Ali Borji. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*, February 2023. doi:10.48550/arXiv.2302.03494. URL https://arxiv.org/abs/2302.03494.

[6] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert, 2023. URL https://arxiv.org/abs/2302.10198. Work in progress. Added results of advanced prompting strategies, e.g., CoT. (19 pages).

[7] Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*, 2023. URL https://doi.org/10.48550/arXiv.2301.08653. arXiv:2301.08653 [cs.SE].

[8] Nigar M. Shafiq Surameery and Mohammed Y. Shakor. Use chat gpt to solve programming bugs. *International Journal of Information Technology and Computer Science (IJITC)*, 3(1):17–22, 2023. doi:10.55529/ijitc.31.17.22. URL https://doi.org/10.55529/ijitc.31.17.22.

[9] Md. Asraful Haque and Shuai Li. The potential use of chatgpt for debugging and bug fixing. *AI and Robotics*, 2(1), 2023. doi:10.4108/airo.v2i1.3276. URL https://doi.org/10.4108/airo.v2i1.3276.

[10] Wei Ma, Shangqing Liu, Wenhan Wang, Qiang Hu, Cen Zhang, Ye Liu, Liming Nie, and Yang Liu. Chatgpt: Understanding code syntax and semantics. *arXiv preprint arXiv:2305.12138*, May 2023. doi:10.48550/arXiv.2305.12138. URL https://arxiv.org/abs/2305.12138.

[11] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine. 2023. Preliminary report on machine translation evaluation with ChatGPT and GPT-4.

[12] Wenxiang Jiao, Wenxuan Wang, Jen-Tse Huang, and Zhaopeng Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, January 2023. doi:10.48550/arXiv.2301.08745. URL https://doi.org/10.48550/arXiv.2301.08745.

[13] Yuan Gao, Ruili Wang, and Feng Hou. How to design translation prompts for chatgpt: An empirical study. *arXiv preprint arXiv:2304.02182*, April 2023. doi:10.48550/arXiv.2304.02182. URL https://arxiv.org/abs/2304.02182.

[14] Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Summit: Iterative text summarization via chatgpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.findings-emnlp.714. URL https://aclanthology.org/2023.findings-emnlp.714/.

[15] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian

Yin, Shiping Yang, and Xiaojun Wan. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, April 2023. doi:https://doi.org/10.48550/arXiv.2304.02554.

[16] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, February 2023. doi:10.48550/arXiv.2302.12095. URL https://arxiv.org/abs/2302.12095.

[17] Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*, April 2023. doi:10.48550/arXiv.2304.11633. URL https://arxiv.org/abs/2304.11633.

[18] Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224, 1961. doi:10.1145/1460690.1460714. URL https://doi.org/10.1145/1460690.1460714.

[19] William A Woods. Progress in natural language understanding: an application to lunar geology. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, pages 441–450, 1973. doi:10.1145/1499586.1499695. URL http://crossmark.crossref.org/dialog/?doi=10.1145%2F1499586.1499695&domain=pdf&date_stamp=1973-06-04.

[20] Ellen M Voorhees. The trec question answering track. *Natural Language Engineering*, 7(4):361–378, 2001. doi:10.1017/S1351324901002789.

[21] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021. doi:10.48550/arXiv.2101.00774. URL https://arxiv.org/abs/2101.00774.

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. doi:10.48550/arXiv.1606.05250.

[23] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016. doi:10.48550/arXiv.1611.09268. URL https://doi.org/10.48550/arXiv.1611.09268.

[24] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, 2015. doi:10.18653/v1/D15-1237.

[25] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi:10.1162/tacl_a_00276.

[26] Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*, 2019. doi:10.48550/arXiv.1912.05200. URL https://arxiv.org/abs/1912.05200. Submitted to LREC 2020.

[27] Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. Neural Arabic question answering. In Wassim El-Hajj, Lamia Hadrich Belguith, Fethi Bougares, Walid Magdy, Imed Zitouni, Nadi Tomeh, Mahmoud El-Haj, and Wajdi Zaghouani, editors, *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy, August 2019. Association for Computational Linguistics. doi:10.18653/v1/W19-4612. URL https://aclanthology.org/W19-4612/.

[28] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1437/.

[29] Danilo Croce, Alexandra Zelenanska, and Roberto Basili. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402. Springer International Publishing, 2018. doi:10.1007/978-3-030-03840-3_29.

[30] Yiming Shao, Dong Wang, Jing Li, et al. A large-scale chinese machine reading comprehension dataset and evaluation platform. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 443–453. ACL, 2018. doi:10.18653/v1/C18-1043.

[31] Seunghyun Lim, Jinhyuk Lee, and Jaewoo Kang. Korquad: Korean question answering dataset. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4689–4694, 2019. doi:10.18653/v1/D19-1356.

[32] Antoine Keraronb et al. Fquad: French question answering dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2020. doi:10.5281/zenodo.3758895.

[33] Hamed Veisi and Hossein Shandi. Development of a persian medical question answering system for diseases and drugs. *Journal of Biomedical Informatics*, 108:103493, 2020. doi:10.1016/j.jbi.2020.103493.

[34] Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. Parsquad: Machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168. IEEE, 2021. doi:10.22133/ijwr.2021.293313.1101.

[35] Arefeh Kazemi, Jamshid Mozafari, and Mohammad Ali Nematbakhsh. Persianqad: The native question answering dataset for the persian language. *IEEE Access*, 10:26045–26057, 2022. doi:10.1109/ACCESS.2022.3157289.

[36] Yasaman Boreshban, Hadi Yousefinasab, and Seyed Ali Mirroshandel. Providing a religious corpus of question answering system in persian. *Journal of Statistical and Data Science*, 6(1):1–15, 2018. doi:10.29252/jsdp.15.1.87. URL https://jsdp.rcisp.ac.ir/article-1-535-en.html.

[37] Jamshid Mozafari, Arefeh Kazemi, Parham Moradi, and Mohammad Ali Nematbakhsh. Peransel: A novel deep neural network-based system for persian question answering. *Computational Intelligence and Neuroscience*, 2022(1):3661286, 2022. doi:10.1155/2022/3661286.

[38] Arefeh Kazemi, Zahra Zojaji, Mahdi Malverdi, Jamshid Mozafari, Fatemeh Ebrahimi, Negin Abadani, Mohammad Reza Varasteh, and Mohammad Ali Nematbakhsh. Farsnewsqa: A deep learning-based question answering system for the persian news articles. *Information Retrieval Journal*, 26(1):3, 2023. doi:10.1007/s10791-023-09417-2.

[39] Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466v1*, 2023. doi:10.48550/arXiv.2302.06466. URL https://arxiv.org/abs/2302.06466v1.

[40] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*, 2023. doi:10.48550/arXiv.2303.07992.

[41] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023. doi:10.48550/arXiv.2301.07597.

[42] Pit Pichappan, M Krishnamurthy, and P Vijayakumar. Analysis of chatgpt as a question-answering tool. *Journal of Digital Information Management*, 21(2):50–60, 2023. doi:10.6025/jdim/2023/21/2/50-61.

[43] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. doi:10.18653/v1/P18-2124.

[44] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016. doi:10.18653/v1/W17-2623.

[45] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020. doi:10.48550/arXiv.2010.01057. URL https://arxiv.org/abs/2010.01057. Submitted on 2 Oct 2020.

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. doi:10.48550/arXiv.1906.08237. URL https://arxiv.org/abs/1906.08237. 1Carnegie Mellon University, 2Google AI Brain Team.

[47] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pretraining by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019. doi:10.48550/arXiv.1907.10529. URL https://arxiv.org/abs/1907.10529. Equal contribution.

[48] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*, 2020. doi:10.48550/arXiv.2001.09694.

[49] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language rep-

resentations. In *International Conference on Learning Representations (ICLR)*, 2020. doi:10.48550/arXiv.1909.11942. URL https://doi.org/10.48550/arXiv.1909.11942.

[50] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *arXiv preprint arXiv:2210.02627*, 2022. doi:10.1162/tacl_a_00530.

[51] Seohyun Back, Akhil Kedia, Sai Chetan Chinthakindi, Haejun Lee, and Jaegul Choo. Learning to generate questions by learning to recover answer-containing sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1516–1529, 2021. doi:10.18653/v1/2021.findings-acl.132.

[52] Souvik Kundu and Hwee Tou Ng. A question-focused multi-factor attention network for question answering. *Department of Computer Science, National University of Singapore*, 2020. doi:10.1609/aaai.v32i1.12065.

[53] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. Densely connected attention propagation for reading comprehension. 2020. doi:10.5555/3327345.3327399.

Hossein Bahak is a software engineer interested in machine learning and natural language processing. He is passionate about working with open-source tools and strives to solve real-world problems using them. In industrial activities, he is involved in developing chatbots and enhancing product images using deep learning. He also conducts research on large language models at the University of Isfahan and has contributed to the development of platforms such as Sajjad and Porsman.

Farzaneh Taheri hold a Master's degree in Software Engineering from Isfahan University, where she also completed her Bachelor's degree with distinction. She is deeply immersed in the fields of Machine Learning and DevOps. Currently, She is thriving as a Machine Learning Engineer.

Arefeh Kazemi received her B.Sc. degree in software engineering and the M.Sc. degree in artificial intelligence from University of Isfahan, Isfahan, Iran, in 2008 and 2010, respectively. She obtained her Ph.D. degree in artificial intelligence field from University of Isfahan in 2017. Her main areas of research interest include Natural Language Processing, Computational Linguistics and Data Mining.

Zahra Zojaji received the B.Sc. degree in software engineering and the M.Sc. degree in artificial intelligence from University of Isfahan, Isfahan, Iran, in 2008 and 2010, respectively. She has completed Ph.D. in artificial intelligence field at Amirkabir University of Technology, Tehran, Iran, in 2017. She is now an assistant professor in Software Engineering branch in University of Isfahan. Her main areas of research interest include Natural Language Processing, Machine Learning, Data Mining and Social Network Analysis.