Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120916

Using ChatGPT and other LLMs in Professional Environments

S. Alaswad[1,*], *T. Kalganova*[2] *and W. S. Awad*[3]

[1]Department of Multimedia Science, College of Information Technology, Ahlia University, Manama, Bahrain
[2]Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge, UB8 3PH, UK
[3]Department of Information Technology, College of Information Technology, Ahlia University, Manama, Bahrain

Received: 19 Jun. 2023, Revised: 20 Jun. 2023, Accepted: 10 Jul. 2023.
Published online: 1 Sep. 2023.

Abstract: Large language models like ChatGPT, Google's Bard, and Microsoft's new Bing, to name a few, are developing rapidly in recent years, becoming very popular in different environments, and supporting a wide range of tasks. A deep look into their outcomes reveals several limitations and challenges that can be further improved. The main challenge of these models is the possibility of generating biased or inaccurate results, since these models rely on large amounts of data with no access to unpublic information. Moreover, these language models need to be properly monitored and trained to prevent generating inappropriate or offensive content and to ensure that they are used ethically and safely. This study investigates the use of ChatGPT and other large language models such as Blender, and BERT in professional environments. It has been found that none of the large language models, including ChatGPT, have been used in unstructured dialogues. Moreover, involving the models in professional environments requires extensive training and monitoring by domain professionals or fine-tuning through API.

Keywords: Artificial Intelligence, Chatbot, ChatGPT, LLM.

1 Introduction

Conversational Large Language Models (LLM) have become extremely advanced in the last two years and capable of performing several tasks in a human-like way. Basically, these models utilize the advancement in Natural Language Processing (NLP) technology to understand, respond and evaluate user inputs [1]. ChatGPT is one of the models that gained great attention as it can produce fluent descriptive response to user queries [2]. Other models like Chatsonic [3], Google's Bard [4] and others are also evolving, providing numerous services to the industry and society.

Dialogues are the key method of interaction with these conversational AI models. Generally, forms of the dialogues (interviews) can be classified into three main types, structured, semi-structured and unstructured [5]. A dialogue is structured when the set and order of questions are prefabricated and cannot be expanded easily during the process of the interview. The semi-structured dialogue is less rigid than the structured form, it has a more open framework of ideas with the ability to refer to preprepared questions during the interview process. In the unstructured dialogues, aims and goals are not clearly defined ahead of the interview, questions are not prepared but come forward during the conversation process. While unstructured type of dialogues is more suitable for gathering qualitative data, the analysis and evaluation of process is more time consuming than the other two types [5].

Overall, Artificial Intelligence (AI) conversational models are considered as a major innovation in AI technology that will transform the way humans interact with computers and other humans. However, when using these models in professional environments like educational, medical, and industrial fields, generated responses need to be supervised to measure its correctness and appropriateness. In this paper we aim at investigating the challenges of using LLMs in the various domains and industries, looking at the types of data and dialogues used in the conversations. We aim to answer the following research questions:

- RQ1: What are the limitations of LLMs when used in professional environments?
- RQ2: What are the datasets used with ChatGPT and other LLMs?
- RQ3: What are the types of dialogues used with ChatGPT?

[*]Corresponding author e-mail: salaswad@ahlia.edu.bh

2 Research Background, Challenges and Motivations

A. Research Background

Modern generative AI models compared with expert systems can create innovative content, while the latter one mainly analyzes and acts on the data available. Expert systems control their contained knowledge bases via an inference engine that employs an if-else rule to generate the data. On the other hand, generative AI employs a discriminator or transformer model to map input information to a concealed trained dataset. Then, a generator model is employed to produce original output in every attempt, including the attempts with similar input prompts. Thus, unsupervised, semi-supervised and supervised learning is performed in accordance with the designated methodology. In accordance with the content produced by the generative AI models, they differ from the content produced by predictive machine learning systems. The latter system is based mainly on discrimination behavior to solve classification or regression problem, while generative AI are capable of discriminating and generating information out from the transformed prompt or input information [1].

In the past two years, large generative models have remarkably increased in which they were able to undertake several types of general tasks in several fields such as question & answering, image generation and many more. Subsequently, several jobs in industry and society are predicted to be substituted with one of these AI models. The most popular and well-known example here is ChatGPT that can efficiently and innovatively convert texts to other texts. Other models are DALLE-2 and Dreamfusion the converts text to images and 3D images respectively. Flamingo, on the other hand, is a model that converts images into text. Models that convert texts to audio or video are Phenaki and AudioLM. Others can convert text to code, scientific text or produce algorithms [1].

Consequently, those previously mentioned models have been used in a wide variety of domains and industries. In the educational domain, for instance, establishing a solid foundation is essential and placing these models as AI teachers need to consider three dimensions to measure the pedagogical ability: 1. speaking as teachers to students, 2. Understanding students' and 3. helping to improve students' understanding [6]. In the construction field, ChatGPT has been used in planning an automated schedule for construction projects. Prompts in natural language were entered into the chat box and consistent output was produced within a few seconds. However, determining the applicability of the responses in relevance to the scope of work needs further inspection [7]. Another example is testing the suitability of ChatGPT in supporting design discipline, in which ChatGPT was reasonably competent to play the role of imaginary designer, user and even products. In addition, it was able to track the dialogues context along multiple prompts of related subjects and to correct the responses through conversations in free form natural language. Nonetheless, increasing the length of dialogue session and simulating user data introduce some difficulties and limitations in the model [8]. Overall, ChatGPT has the potential to transform the way we interact with machines and each other in a wide variety of domains and industries.

Looking at the datasets used with LLMs can provide an insight of the dialogue types used. Tack and Piech used two datasets in measuring the pedagogical ability of Blender and GPT-3. The datasets are 1) The Educational Uptake Dataset and 2) The Teacher-Student Chatroom Corpus (TSCC). The two datasets have a total of 6,685 dialogic pairs with the utterances annotated to identify the organization of the conversation. Since the student-teacher dialogue is open-ended with several ways in which the teacher agent can respond to the student, the dialogues are classified as semi-structured type of dialogue [6]. In mathematics, GHOSTS (a collection of multiple prompts datasets) is created covering mathematics of graduate level and was used to inspect the mathematical capabilities of LLMs. The collection is divided into six sub datasets: Grad-Text, Holes-in-Proofs, Olympiad-Problem-Solving, Symbolic-Integration, MATH, Search-Engine-Aspects. The total number of used prompts was 728 and the output of ChatGPT was evaluated by domain professionals [9].

In machine translation, Jiao et al. [10] sampled 50 sentences from each of the testing sets: General Language, Biomedical, Reddit and Common voice, as ChatGPT can't respond to large batches of prompts.

Prieto, Mengiste and De Soto [7] designed an experiment to evaluate ChatGPT in scheduling construction projects. The experiment is based on a textual scenario carried out by expert participants allowing them to challenge, modify and evaluate the model response [7]. The same approach (Textual Scenario) was used by Kocaballi in assessing the design capabilities of ChatGPT, where a designer (researcher) interacts with the model for performing different design activities [8]. In the textual scenario, the human starts with providing the scope of the domain, then prompts are designed (ChatGPT/human generated). Based on the response provided by the model, there might be some variations on the prompts or sometimes the domain by the human. At the end the appropriateness of outputs is evaluated by humans and some defined metrics [8]. Similarly, software architecture with ChatGPT has three phases. In phase 1, text-based architecture story is developed and fed to ChatGPT. Next in phase 2, collaborative architecting with Software engineers who analyze, synthesize, and evaluate the process. In the last phase, empirical validation by software engineering team is conducted using surveys [10].

Textbooks were also used to form datasets in some fields while exploring ChatGPT. In software testing, 40 questions were extracted from five chapters from a software testing textbook and ChatGPT was used to provide three responses for each question. The result is a dataset of the size 120. Some questions are asking for a code or explaining a concept or both. Other questions are asking for a set of instructions to accomplish a specific task [11]. In the Algebra topic area, questions were obtained from OpenStax Elementary Algebra and Intermediate Algebra textbooks [12].

For measuring language understanding of ChatGPT, GLUE dataset with several NLP tasks were used. The tasks are either classification tasks (linguistic acceptability, sentiment analysis, paraphrase, question paraphrase, textual entailment, and question-answer entailment) or regression task (textual similarity) [2].

In the medical field, questions of Radiation Oncology In-Training Exam (TXIT) and 2022 Red Journal gray zone cases were used to measure the performance of ChatGPT-3.5 and ChatGPT-4. Six categories in radiation oncology are covered by the TXIT exam which contains 300 questions, 7 questions were eliminated as they require access to image information. For the 2022 collection of the Red Journal gray zone cases, ChatGPT-4 was only used as a radiation oncology expert to answer a total of 15 cases [13].

Table 1 looks at the datasets and number of dialogues used with ChatGPT. In some cases, like construction and design fields, textual scenarios by experts from the domain were used to examine the capabilities of ChatGPT. Looking at the dialogue types, scripted and semi-scripted dialogues are mostly used to form the dialogues, unstructured dialogues have never been used with ChatGPT in any of the investigated domains.

Table 1: Datasets used with ChatGPT in different domains

Authors, Year	**Domain**	**Dataset**	**Dialogue Size**	**Dialogue Type**
A. Tack and C. Piech, 2022 [2]	Language	TSCC	4439	semi-structured
	Mathematics	Uptake	2246	
S. Frieder et al., 2023 [14]	Mathematics	GHOSTS	728	structured
W. Jiao et al., 2023 [15]	Machine Translation	General Language	1012	structured
		Biomedical	373	
		Reddit	2373	
		Common voice	5609	
S. A. Prieto, E. T. Mengiste, and B. G. De Soto, 2023 [7]	Construction	Textual Scenario	6 cases	semi-structured
S. Jalil et al., 2023 [11]	Software Testing	Textbook	120	semi-structured
Z. A. Pardos and S. Bhandari, 2023 [12]	Algebra	Textbook	-	semi-structured
A. B. Kocaballi, 2023 [8]	Design	Textual Scenario	1 case	semi-structured
Q. Zhong et al., 2023 [2]	Language Understanding	GLUE Dataset	425	structured
A. Ahmad et al., 2023 [10]	Architecture-centric Software Engineering	Textual Scenario	-	semi-structured
Y. Huang *et al.*, 2023 [13]	Medicine	ACR TXIT exam 2022 Red Journal gray zone cases	300 14 cases	semi-structured

B. Challenges

Though ChatGPT and other language models have demonstrated their capabilities in various professional environments, they also face some challenges and limitations. Mainly, these language models rely on the data they were trained on, which may contain biases or inaccuracies that result in biased and inaccurate responses. And accuracy and fairness are major concerns in professional settings such as medical, educational or legal domains [11]. Furthermore, as ChatGPT and other models may not be able to interpret complex and ambiguous queries, this might lead to incomplete or misleading response due to the lack of contextual understanding [8]. Moreover, handling confidential or sensitive information is also challenging as these models are based on machine learning, which requires human judgment and ethical considerations. Lastly, these models might be exploited in professional environments to generate fraudulent documents and spread misinformation [1].

In the following sub sections, we address four main challenges of ChatGPT and other language models that are critical to ensure responsible and ethical use of the models in professional environments.

1) Understanding

We can gain insight into the understanding ability of AI models from the educational field into how the teacher responds to a student. In a pilot study, AI teachers (Blender and GPT- 3) were run in parallel to human teachers in mathematics and language educational dialogues. When it comes to understanding the student, Blender outperforms GPT-3 and the actual teacher on this specific pedagogical dimension. Oppositely, GPT-3 was continually behind Blender and human teachers in all measured abilities. Moreover, when it comes to helping the student, both models were behind the real teachers' performance [2].

Another study concentrates on measuring the mathematical capabilities of ChatGPT, where a dataset is created with a total of 728 prompts and manual rating of output by human experts. Results show that ChatGPT is well-below an average mathematic graduate student, however ChatGPT can demonstrate a clear understanding of the question but not able provide a solution in a complete correct manner [14].

Shifting to another field, ChatGPT has been used to play different roles in the design field: designer, user, and product. The agent was able to understand the prompts in natural language and to generate different designer personas, interview questions, interview simulation and ideation on a project titled "Designing a voice assistant for the health and wellbeing of people working from home". The whole study was conducted in one session to maintain project context. However, there were some drawbacks related to forgotten information, incomplete responses, and lack in the diversity of responses. These limitations were referenced in the documentation provided by OpenAI, as the model can understand up to 4000 tokens (or 3000 words), no more information is stored beyond this limit [8].

Zhong et al [2]. explored the language understanding of ChatGPT by using the GLUE benchmark and the seven datasets mentioned previously. ChatGPT failed to handle paraphrase and similarity tasks but exceeds BERT models in inference tasks. Researchers found that advanced prompting strategies can have significant improvement on ChatGPT performance. In the medical field, providing in-context information to ChatGPT enables it to help clinicians to understand the latest guidelines. Hence, updated treatment was provided to patients as per the updated guidelines [13].

2) Responses

The second challenge we would like to address here is the appropriateness of the responses generated by the AI models and their effectiveness to accomplish a specific task. Referring to the dimensions of the pilot study of AI teachers, the first dimension was to assess if the model is responding to student as a teacher. In terms of pedagogical ability, Blender (as an AI teacher) can expand on student's utterance and score higher evaluation percentage of positive responses. In contrast, ChatGPT was behind Blender and human teachers in this dimension, and it is important to mention that not all human teacher' responses were evaluated positively [6].

ChatGPT was also used as a machine translator in a preliminary study that focuses on three main aspects: 1. Translation Prompts, 2. Multilingual Translation, and 3. Translation Robustness. As the syntax of prompts in stating the source and target languages affects the translation quality, translation prompts were obtained by ChatGPT itself. Looking into the second aspect, multilingual translation, ChatGPT was able to compete with other commercial products (Google Translate, DeepL Translate, and Tencent TranSmart) especially in languages with high resources like European languages. For low-resource and distant languages, a strategy called pivot prompting was utilized to enhance the overall performance, where the source sentence is translated to a pivot high-resource language prior to the target language. And regarding the translation robustness aspect, ChatGPT was examined in domain-specific sentences like biomedical abstracts and Reddit comments, where it shows a performance that is lower than other commercial systems [15].

Another preliminary study was conducted to investigate the usage of ChatGPT for the scheduling of construction projects. The study involves a number of participants interacting with GPT to develop an automated construction schedule for a simple project based on detailed input prompts in natural language description. Participants evaluated the whole process, and they were impressed by the coherence, reasonability, and the speed at which output was generated. Further inspection into the responses revealed that some of the proposed tasks are not within the scope of work, whilst other tasks weren't considered at all. This limitation is due to the fact that ChatGPT needs to be trained on specific construction purposes to be aware of the type and sequence of tasks to be scheduled [7].

For the correctness of responses in the study of mathematical capabilities of ChatGPT, it was found that even if ChatGPT can understand the question, it still fails to provide consistent proofs and high-quality calculations in some math respects [14]. In another study examining the software testing capabilities of ChatGPT, 40 questions were taken from a popular software testing curriculum, in which ChatGPT correctly or partially answered an approximate 44% of the cases. The model was also able to correctly answer or partially an approximate of 57% explanation questions. The study concludes that asking the questions and their sub-questions in a single chat (shared context) improves the rate of getting correct answers rather than separate context [11].

Now for the appropriateness of responses in the design domain, the knowledge of ChatGPT as a large language model is still considered debatable in terms of accuracy. Hence, expert people from the domain are required for fact-checking and more reliable use. It was concluded that AI models fit more in content summarization, solutions ideation and conversations simulation [8]. Another use case is the capability of ChatGPT to architect and collaborate with humans in architecture-centric software engineering (ACSE). Primary results denote that although ChatGPT can simulate the role of architect, support and lead ACSE, there is still a need for human supervision in decision making and collaborative architecting [10]. In sports science, psychology, and MCQ-based exams, ChatGPT's answers have also failed to meet the passing grade. However, GPT-4-based version passed most MCQ-based exams with a comparable performance to human subject [16-17].

Surameery and Shakor explored the capabilities of ChatGPT in solving programming problems and how it can assist in debugging, bug prediction and explanation. ChatGPT demonstrated understanding and analysis abilities of code snippets. However, ChatGPT still can't be considered as a perfect tool, since the quality of the responses highly depends on the system design and training dataset. Authors recommended using alternative debugging tools to validate ChatGPT outputs [18].

In an investigation of ChatGPT in healthcare education, research and practice, Sallam recorded several limitations including transparency, legal issues, risks of bias, hallucination and infodemics. The author concluded that the adoption of this model has to be monitored carefully and effectively [19]. In addition to all previously mentioned use cases of ChatGPT, the last use case to be mentioned here is ChatGPT as an oncologist. The model demonstrates a good knowledge in a number of topics such as pediatrics, biology, physics and others, but lacks proficiency in topics like bone & soft tissue, gynecology, dosimetry and brachytherapy, along with in-depth questions from clinical trials [13].

Table 2 evaluates ChatGPT in terms of ability of understanding and appropriateness of responses. In most of the domains, it was clear that although ChatGPT can understand the prompts correctly, it usually fails to provide correct and trustable outputs. Later in Table 3, it will be shown that automated methods or humans or both were used to evaluate the understanding and responses of the models based on the type of dialogues used in the datasets.

Table 2: Evaluating the ability of understanding and appropriateness of responses of ChatGPT in different domains

Authors, Year	**Domain**	**Uses**	**Understanding**	**Responses**	**ChatGPT Version**
A. Tack and C. Piech, 2022 [2]	Language & Mathematics	Observe how Blender and GPT-3 respond to a student, and compare their responses in terms of pedagogical ability	×	×	GPT-3
S. Frieder et al., 2023 [14]	Mathematics	Measure ChatGPT mathematical capabilities	×	×	GPT-3
W. Jiao et al., 2023 [15]	Machine Translation	Perform machine translation tasks	✓	×	GPT-3.5
S. A. Prieto, E. T. Mengiste, and B. G. De Soto, 2023 [7]	Construction	Evaluate ChatGPT to assist in developing an automated construction schedule based on natural language prompts	✓	×	GPT-3.5
S. Jalil et al., 2023 [11]	Software Testing	ChatGPT provided 3 responses for 40 questions from five chapters of a popular software testing textbook	✓	×	GPT-3.5
Z. A. Pardos and S. Bhandari, 2023 [12]	Algebra	Investigate if ChatGPT generated hints can be beneficial to algebra learning	✓	×	GPT-3.5
A. B. Kocaballi, 2023 [8]	Design	Understand the capabilities, limitations, and overall suitability of ChatGPT as a large language model to support the design process	×	×	GPT-3.5
Q. Zhong et al., 2023 [2]	Language Understanding	Measure the understanding ability of ChatGPT by	✓	✓	GPT-3.5

		evaluating it on the most popular GLUE benchmark, and comparing it with 4 representative fine-tuned BERT-style models			
A. Ahmad et al., 2023 [10]	Architecture-centric Software Engineering	ChatGPT collaborated with a novice software for architectural analysis, synthesis, and evaluation of a services-driven software application	✓	✗	GPT-3.5
Y. Huang et al., 2023 [13]	Medicine	Benchmark the performance of ChatGPT-3.5 and ChatGPT-4 on a radiation oncology exam and some other cases.	✓	✓	GPT-3.5 GPT-4
A. Szabo, 2023 [16]	Sports Science & Psychology	Test ChatGPT's information accuracy on exercise addiction	✓	✗	GPT-3.5
P. M. Newton, M. Xiromeriti, and U. Kingdom, 2023 [17]	MCQ-based exams across subjects	Review the performance of ChatGPT on MCQ-based exams	✓	✗	GPT-3 GPT-3.5 GPT-4
N. M. S. Surameery and M. Y. Shakor, 2023 [18]	Computer Programing	Use ChatGPT to solve programming bugs	✓	✗	GPT-3.5
M. Sallam, 2023 [19]	Healthcare	Investigate the utility of ChatGPT in healthcare education, research, and practice.	✓	✗	GPT-3

3) Assessment

As AI continues to develop, so does the need for effective evaluation and assessment of AI models. AI model assessment is crucial to ensure that AI models are accurate, reliable, and ethical. An insight in this regard can be gained from previous work into human teachers' evaluation as the educational field is rich in assessment methods for measuring teacher effectiveness. Methods include self-reports, interviews, in-class peer reviewing, students' surveys and students' outcomes measurements. Nevertheless, some of these methods are not easily applicable to AI teachers and would be difficult to implement. However, methods like AI teachers' observation, evaluation of human surveys and students' achievement can be systematically automated [6].

There are several factors that need to be considered when assessing AI models. Training and testing dataset, sampling techniques to ensure using unbiased data, evaluation metrics, social impact, and ethical considerations [9]. In the studies measuring the ability in teaching, establishing good assessment methods of AI teachers is essential and hard at the same time. Furthermore, there is no universal solution to evaluate teaching effectiveness and ability. Consequently, two datasets: Language (TSCC) and Mathematics (Uptake) were used to quantitatively compare AI responses with the responses of human teachers to measure the pedagogical ability [6].

In other studies, to ensure reliable use of the model, human participants with sufficient domain knowledge should be involved, and the results obtained are to be evaluated in terms of accuracy, efficiency, clarity, coherence, reliability, relevance, consistency, scalability, and adaptability [7]. A good example here is when ChatGPT used in the design field, an experienced HCI designer and researcher in the areas of conversational user interfaces, human-AI interaction, and digital health was involved to analyze the appropriateness of the responses based on qualitative assessment [8]. In the medical field, specialized domain experts verified all answers and recommendations by ChatGPT by applying sophisticated clinical reasoning and using qualitative and semiquantitative evaluation methods [13].

Although results from machine translation study can be calculated using automatic metrics like: BLEU score ChrF++ and TER, since scripted dialogues can be utilized in this domain, human evaluation can always reflect more translation aspects (nativeness as an example) and provide better understanding to compare ChatGPT results with other commercial systems.

Table 3 shows interaction modality, dialogue types, assessment metrics, and human involvement in assessing ChatGPT in different domains.

Table 3: Assessing ChatGPT in different domains

Authors, Year	**Domain**	**Interaction Modality**	**Metric Name**	**Human?**
A. Tack and C. Piech, 2022 [6]	Language	Text-to-text	BLEU F1	Yes
	Mathematics		MAUVE	
R. Gozalo-Brizuela and E. C. Garrido-Merchan, 2023 [1]	Image Generation	Text-to-image	BERT	No
W. Jiao et al., 2023 [15]	Machine Translation	Text-to-text	BLEU score ChrF++ TER	No
S. A. Prieto, E. T. Mengiste, and B. G. De Soto, 2023 [7]	Construction	Text-to-text	-	Yes
A. B. Kocaballi, 2023 [8]	Design	Text-to-text	-	Yes
Q. Zhong et al., 2023 [2]	Language Understanding	Text-to-text	GLUE Benchmark	No
A. Ahmad et al., 2023 [10]	Architecture-centric Software Engineering	Text-to-text	SAAM Or ATAM	Yes
Y. Huang et al., 2023 [13]	Medicine	Text-to-text Image-to-text	-	Yes

4) Hallucination

Hallucination in LLMs is the phenomenon in which the models may produce incorrect or fictional responses that are not supported by the training dataset. It's considered as one of the significant challenges in this field despite the extraordinary performance in the understanding and NLP tasks. This issue supports the concerns about the accuracy, reliability, and ethical aspects of employing these AI models in research and professional environments. In this section we discuss the causes and implications of hallucination in LLMs to mitigate its impact and to ensure accountable responses of the models [20-21]. Lacking ground truth data and up-to-date source of information is one of the main roots of hallucination in some LLMs. For instance, ChatGPT-3.5 has no awareness of the events that occurred after September 2021, since the datasets used in the training phase were collected before that date. This issue was addressed in other LLMs such as Microsoft's new Bing [22] and Google's Bard [4] in which they process information retrieved from the internet in real-time. Accordingly, to ensure providing accuracy context to the users, LLM should be able to select credible sources of information [23]. Other factors leading to hallucination in LLMs is the inference mechanisms used, which can be enhanced using prompt engineering techniques such as UPRISE [24]. SelfCheckGPT is one of the approaches developed for detecting intrinsic hallucinations in LLM as it can verify the trust- worthiness of the output generated at sentence and passage levels [25].

The hallucination in ChatGPT can be classified into two categories: intrinsic hallucinations and extrinsic hallucinations. The first type of hallucination is when the model generates an output that contradicts the input content, while the output in the second type can't be contradicted neither supported by the input content [26]. Machine Translation is one of the fields that suffers from hallucinations in non-English tasks and multilingual translation models trained on low-resource language pairs [27-28]. Besides, causal reasoning hallucinations are serious in ChatGPT and it usually assumes causal relations between unrelated events [29].

Table 4 illustrates ChatGPT hallucination in different domains while accomplishing domain-specific tasks.

Table 4: Hallucination of ChatGPT in different domains

Authors, Year	**Domain**	**Uses**
S. Ott et al., 2023 [20]	• Scientific/Medical • General Domain	• Evaluate logical reasoning

	• Math		
S. Cahyawijaya et al., 2023 [26]	• Question Answering • Reasoning, Summarization • Machine Translation	• Automatic Post-Editing • Sentiment Analysis	• Evaluate multitask, multilingual and multimodal aspects
D. Cheng et al., 2023 [24]	• Reading Comprehension • Closed-Book QA • Paraphrase Detection • Natural Language Inference • Sentiment Analysis	• Commonsense Reasoning • Coreference Resolution • Structure To Text • Summarization	• Achieve universality of the prompt retriever
L. Ding, Q. Zhong, and L. Shen, 2023 [27]	• Machine Translation		• Machine translation tasks
N. M. Guerreiro et al., 2023 [28]	• Machine Translation		• Machine translation tasks
K. Yang and F. Menczer, 2023 [23]	• News Outlet Credibility		• Rate news outlet credibility
H. Gilbert et al., 2023 [21]	• Semantic Compression		• Quantify the capability of LLMs to compress text and code
J. Gao, X. Ding, B. Qin, and T. Liu, 2023 [29]	• General Domain		• Evaluate causal reasoning

C. Research Motivations

Several motivating factors encourage the utilization of ChatGPT and other LLMs for research purposes in professional environments. Initially, the massive size of datasets that is used to train the models, which encompass a wide range of domains and resources that enable researcher to have diverse views and perspectives. In addition to that, the models facilitate efficient data analysis and extraction due to the advanced NLP capabilities. Hence, it provides professionals with powerful means to accelerate research process leading to strong conclusions, pattern identification and hidden correlations reveal. Moreover, experts can use the models to generate hypotheses and ideas, or to refine them by inputting specific prompts and getting feedback in real-time. Furthermore, as the models have been trained on multilingual repositories, this extends the research scope and facilitates universal research collaboration. The last motivating factor to be mentioned here is the ability of LLMs to be adapted and fine-tuned for targeting domain-specific tasks.

Nonetheless, involving the models in professional frameworks exposes numerous limitations related to general issues such as ethical concerns, resources law-compliance, information correctness and accuracy. Evaluation methods of generated responses is another issue, where there is a need to establish unified criteria and automated assessment process. Other research concerns are associated with the datasets used to examine the capabilities of the AI models and the types of dialogues used for interaction. This is considered as another key dimension which has a principal impact on the overall performance of the models and the generated output.

To sum up, the motivations for engaging ChatGPT and other LLMs in professional research environments are numerous. Furthermore, by utilizing the power of LLMs, new insights, innovations and contributions can be made in every single field.

3 Current Research Status

The development of AI-powered tools is an expanding field, with continuous introduction and updates to new and existing tools. Therefore, it's possible that when this paper gets published, advancements or newer features of these tools may be released. Nevertheless, we illustrate in this section a few currently available AI tools, particularly intended for professional environments, organizing them into four main categories based on the main task accomplished by the tool. Figure 1 illustrates the four main categories of most recent AI tools.

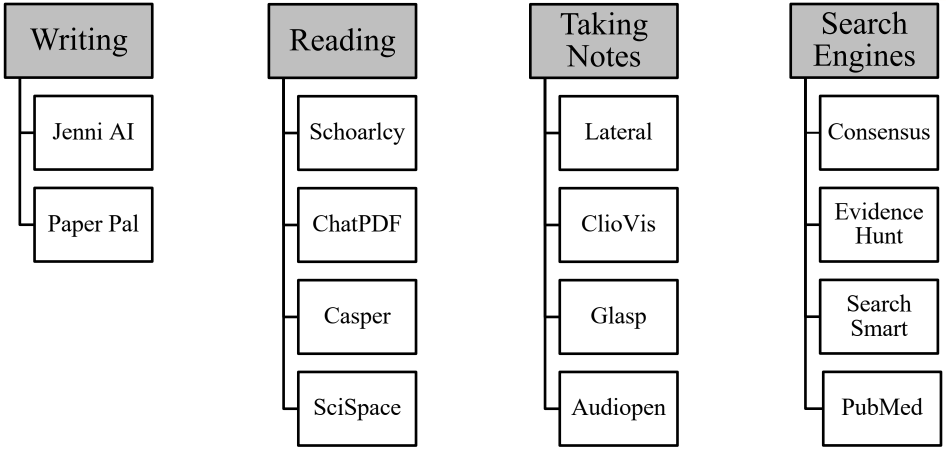

Fig. 1: Recently appeared AI-powered tools classified based on the primary accomplished task

A. Writing Tools

In professional and academic writing there is an increasing demand for intelligent writing tools with advanced writing abilities to meet the high writing standards in the academic field. Jenni AI and Paperpal are two writing tools that play the role of intelligent writing assistant by leveraging the capabilities of AI and machine learning. Both tools apply NLP algorithms and ML techniques to understand and analyze text-based prompts, enabling professionals to polish their writing skills and to produce high-quality contents in their respective domains. For instance, grammar, punctuation, and style suggestions available in Jenni AI, enhance the overall readability and coherence of the writing [30]. While Paperpal focuses mainly on scholarly and scientific writing for students, professionals, and researchers. The tool has an MS Word plugin and features to manage citations, check grammar, detect plagiarism and ensure the academic integrity of the work [31].

B. Reading Tools

Another important demand in the research context is related to the reading experience. Innovative tools are needed to support effective information extraction and optimize the reading process for respective people. Scholarcy, ChatPDF, Casper, and SciSpace are four reading tools to support professionals in this aspect. Scholarcy is a reading assistant that grants access to a massive amount of research papers and scholarly articles, with powerful search capabilities and the ability to summarize research papers with unfamiliar terms hyperlinked to Wikipedia entries [32]. ChatPDF, on the other hand, allows the researcher to upload a normal static pdf document and transform it into an interactive one. This is achieved by enabling the researcher to ask questions and the tool provides answers obtained from the uploaded document. The tool also facilitates research collaboration among users as they can work and share research insights together [33].

Casper is a Chrome extension that summarizes research papers within browser. It allows readers to process information efficiently, enhances reading speed and helps to brainstorm ideas [34]. Lastly here is SciSpace which is another resourceful tool designed for scientific literature. It enables researchers to have a deeper understanding of complex research articles by asking the tool copilot to explain difficult passages [35].

C. Taking Notes Tools

Tools in this section are intended to improve the organization, analysis, and retrieval of information in academic research. Example tools are Lateral, ClioVis, Glasp, and Audiopen. The research tool, Lateral, is an application that helps to discover common themes and connections within multiple research papers in a few minutes. CliVis, alternatively, is purposely designed for people in history and humanities research. It has features to support generating timeline and mind map to enable users to visualize and organize historical events and sources [36]. Glasp supports multimodality allowing users to capture, annotate, tag and organize the different types of media files. Thus, they can easily retrieve any specific piece of information later [37]. Audiopen tool uses smart pens with capabilities to record audio and synchronize them with written notes. Users can later revisit certain points in the interview, meetings, or lectures by clicking on the written notes only [38].

D. Search Engine Tools

The tools listed in this section are search tools that can answer questions with references to published paper. Consensus, for example, is a platform to facilitate collaborative discussions with evaluations and ranking options [39]. Evidence

Hunt cites scientific literature, databases, and search repositories for clinical type of questions by utilizing advanced search algorithm along with the NLP techniques [40]. Search Smart searches for databases based on the researcher's search history and interest. Hence, it enhances the research and exploration process [41]. PubMed is another research tool that supports searching and retrieving literature related to biomedicine and life sciences. The database includes biomedical literature resources of more than 35 million citations and abstracts, with links to the full text journal articles. [42].

4 Conclusions

ChatGPT and other LLMs revolutionized many industries including healthcare, construction, education, design, machine translation and many more. However, challenges like 1. ability of models' understanding, 2. appropriateness of generated responses, 3. assessments and evaluation metrics, and 4. hallucination aspects are considered essential limitations that need to be considered when involving models in professional environments. This paper discussed the previously mentioned challenges by investigating the use of ChatGPT and other LLMs in different types of professional environments. Additionally, we looked into the datasets and types of dialogues used with the models to accomplish domain-specific tasks. The paper discovered that unstructured dialogues have never been used with any of the LLMs including ChatGPT. Furthermore, the models are not sufficiently trained for professional environments unless they are monitored by domain professionals or fine-tuned through API.

Conflict of interest

The authors declare that there is no conflict regarding the publication of this paper.

References

[1] R. Gozalo-Brizuela and E. C. Garrido-Merchan, ChatGPT is not all you need. A State of the Art Review of large Generative AI models, *arXiv:2301.04655v1*, (2023).

[2] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT, *arXiv:2302.10198*, (2023).

[3] Writesonic, ChatGPT Alternative Built With Superpowers - Chatsonic (Now GPT-4 Powered), (2023). [Online]. Available: https://writesonic.com/. [Accessed: 26-May-2023].

[4] S. Pichai, An important next step on our AI journey, (2023). [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/. [Accessed: 26-May-2023].

[5] L. Decher, *Qualitative Interviews. Conducting Interviews as a Means of Qualitative Study. Main Types of Interviews*. GRIN Verlag, (2016).

[6] A.Tack and C. Piech, The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues, *arXiv:2205.07540v1*, (2022).

[7] S. A. Prieto, E. T. Mengiste, and B. G. De Soto, Investigating the use of ChatGPT for the scheduling of construction projects, *arXiv:2302.02805*, 1–14, (2023).

[8] A.B. Kocaballi, Conversational AI-Powered Design: ChatGPT as Designer, User, and Product, *arXiv:2302.07406*, (2023).

[9] C. Leiter, R. Zhang, Y. Chen, J. Belouadi, D. Larionov, and V. Fresen, ChatGPT: A Meta-Analysis after 2.5 Months, *arXiv:2302.13795v1*, (2023).

[10] A.Ahmad, M. Waseem, P. Liang, M. Fehmideh, M. S. Aktar, and T. Mikkonen, *Towards Human-Bot Collaborative Software Architecting with ChatGPT*, in Proc. the 27th International Conference on Evaluation and Assessment in Software Engineering, 279–285, (2023).

[11] S. Jalil, S. Rafi, T. D. LaToza, K. Moran, and W. Lam, *ChatGPT and Software Testing Education: Promises & Perils*, in 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 4130–4137, (2023).

[12] Z. A. Pardos and S. Bhandari, Learning gain differences between ChatGPT and human tutor generated algebra hints, *arXiv: 2302.06871.v1*, (2023).

[13] Y. Huang *et al.*, Benchmarking ChatGPT-4 on ACR Radiation Oncology In-Training Exam (TXIT): Potentials and Challenges for AI-Assisted Medical Education and Decision Making in Radiation Oncology, *arXiv:*

2304.11957, (2023).

[14] S. Frieder *et al.*, Mathematical Capabilities of ChatGPT, *arXiv:2301.13867v1*, (2023).

[15] W. Jiao, W. Wang, J. H. Xing, and W. Zhaopeng, Is ChatGPT A Good Translator? A Preliminary Study, *arXiv:2301.08745v2*, (2023).

[16] A.Szabo, ChatGPT a breakthrough in science and education: Can it fail a test?, *OSF Prepr.*, (2023).

[17] P. M. Newton, M. Xiromeriti, and U. Kingdom, ChatGPT Performance on MCQ Exams in Higher Education. A Pragmatic Scoping Review, *EdArXiv doi:10.35542/osf.io/sytu3*, (2023).

[18] N. M. S. Surameery and M. Y. Shakor, Use Chat GPT to Solve Programming Bugs, *IJITC*, vol. 3, no. 1, (2023).

[19] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, *Healthcare*, vol. 11, no. 6, (2023).

[20] S. Ott *et al.*, ThoughtSource: A central hub for large language model reasoning data, *arXiv: 2301.11596*, (2023).

[21] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-smith, and J. White, Semantic Compression With Large Language Models, *arXiv: 2304.12512*, (2023).

[22] Bing, Introducing the new Bing, (2023). [Online]. Available: https://www.bing.com/new. [Accessed: 26-May-2023].

[23] K. Yang and F. Menczer, Large language models can rate news outlet credibility, *arXiv: 2304.00228*, (2023).

[24] D. Cheng, S. Huang, J. Bi, Y. Zhan, and J. Liu, UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation, *arXiv: 2303.08518*, (2023).

[25] P. Manakul, A. Liusie, and M. J. F. Gales, SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, *arXiv: 2303.08896*, (2023).

[26] S. Cahyawijaya, N. Lee, W. Dai, D. Su, and B. Wilie, A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity, *arXiv: 2302.04023*, (2023).

[27] L. Ding, Q. Zhong, and L. Shen, Towards Making the Most of ChatGPT for Machine Translation, *arXiv: 2303.13780*, (2023).

[28] N. M. Guerreiro *et al.*, Hallucinations in Large Multilingual Translation Models, *arXiv: 2303.16104*, (2023).

[29] J. Gao, X. Ding, B. Qin, and T. Liu, Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation., *arXiv: 2305.07375*, (2023).

[30] Jenni, Supercharge Your Next Research Paper, (2022). [Online]. Available: https://jenni.ai/. [Accessed: 26-May-2023].

[31] Paperpal, Experience the future of academic writing, (2023). [Online]. Available: https://paperpal.com/. [Accessed: 26-May-2023].

[32] Scholarcy, The AI-powered article summarizer, (2023). [Online]. Available: https://www.scholarcy.com/. [Accessed: 26-May-2023].

[33] ChatPDF, Upload and Chat PDF Analyze .pdf files with AI, (2023). [Online]. Available: https://pdf.chat/. [Accessed: 26-May-2023].

[34] O. HQ, CasperAI Demo, (2023). [Online]. Available: https://casperai.xyz/. [Accessed: 26-May-2023].

[35] SCISPACE, Simplify scientific knowledge discovery, (2023). [Online]. Available: https://typeset.io/discover/. [Accessed: 26-May-2023].

[36] ClioVis, ClioVis visualizing connections, (2023). [Online]. Available: https://cliovis.com/. [Accessed: 26-May-2023].

[37] Glasp, Social Web Highlighter, (2023). [Online]. Available: https://glasp.co/. [Accessed: 26-May-2023].

[38] AudioPen, The easiest way to convert messy thoughts into clear text, (2023). [Online]. Available: https://audiopen.ai/#! [Accessed: 26-May-2023].

[39] CONSENSUS, Evidence-Based Answers, Faster, (2023). [Online]. Available: https://www.consensus.app/. [Accessed: 26-May-2023].

[40] EvidenceHunt, EvidenceHunt allows you to search for clinical evidence in a quick and effective way, (2023). [Online]. Available: https://evidencehunt.com/. [Accessed: 26-May-2023].

[41] SEARCHSMART, Compare 95 academic databases and their search systems, (2023). [Online]. Available: https://www.searchsmart.org/. [Accessed: 26-May-2023].

[42] NLM, PubMed, (2023). [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/. [Accessed: 12-Jun-2023].