Meta-Radiology 1 (2023) 100017

Contents lists available at ScienceDirect

Meta-Radiology

journal homepage: www.keaipublishing.com/en/journals/meta-radiology

Summary of ChatGPT-Related research and perspective towards the future of large language models

Yiheng Liu [a,1], Tianle Han [a,1], Siyuan Ma [a], Jiayue Zhang [a], Yuanyuan Yang [a], Jiaming Tian [a], Hao He [a], Antong Li [b], Mengshen He [a], Zhengliang Liu [c], Zihao Wu [c], Lin Zhao [c], Dajiang Zhu [d], Xiang Li [e], Ning Qiang [a], Dingang Shen [f,g,h], Tianming Liu [c], Bao Ge [a,*]

[a] School of Physics and Information Technology, Shaanxi Normal University, Xi'an, 710119, Shaanxi, China
[b] School of Life and Technology Biomedical-Engineering, Xi'an Jiaotong University, Xi'an, 710119, Shaanxi, China
[c] School of Computing, The University of Georgia, Athens, 30602, USA
[d] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, 76019, USA
[e] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, 02115, USA
[f] School of Biomedical Engineering, ShanghaiTech University, Shanghai, 201210, China
[g] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, 200230, China
[h] Shanghai Clinical Research and Trial Center, Shanghai, 201210, China

A B S T R A C T

This paper presents a comprehensive survey of ChatGPT-related (GPT-3.5 and GPT-4) research, state-of-the-art large language models (LLM) from the GPT series, and their prospective applications across diverse domains. Indeed, key innovations such as large-scale pre-training that captures knowledge across the entire world wide web, instruction fine-tuning and Reinforcement Learning from Human Feedback (RLHF) have played significant roles in enhancing LLMs' adaptability and performance. We performed an in-depth analysis of 194 relevant papers on arXiv, encompassing trend analysis, word cloud representation, and distribution analysis across various application domains. The findings reveal a significant and increasing interest in ChatGPT-related research, predominantly centered on direct natural language processing applications, while also demonstrating considerable potential in areas ranging from education and history to mathematics, medicine, and physics. This study endeavors to furnish insights into ChatGPT's capabilities, potential implications, ethical concerns, and offer direction for future advancements in this field.

1. Introduction

Recent advances in natural language processing (NLP) have led to the development of powerful language models such as the Generative Pre-trained Transformer (GPT) series,[1,2,3–5] including large language models (LLM) such as ChatGPT (GPT-3.5 and GPT-4).[6] These models are pre-trained on vast amounts of text data and have demonstrated exceptional performance in a wide range of NLP tasks, including language translation, text summarization, and question-answering. In particular, the ChatGPT model has demonstrated its potential in various fields, including education, healthcare, reasoning, text generation, human-machine interaction, and scientific research.

A key milestone of LLM development is InstructGPT,[2] a framework that allows for instruction fine-tuning of a pre-trained language model based on Reinforcement Learning from Human Feedback (RLHF).[7,2] This framework enables an LLM to adapt to a wide range of NLP tasks, making it highly versatile and flexible by leveraging human feedback. RLHF enables the model to align with human preferences and human values, which significantly improves from large language models that are solely trained text corpora through unsupervised pre-training. ChatGPT is a successor to InstructGPT. Since its release in December 2022, ChatGPT has been equipped with these advanced developments, leading to impressive performances in various downstream NLP tasks such as reasoning and generalized text generation. These unprecedented NLP capabilities spur applications in diverse domains such as education, healthcare, human-machine interaction, medicine and scientific research. ChatGPT has received widespread attention and interest, leading to an increasing number of applications and research that harness its exceeding potential.

The open release of the multi-modal GPT-4 model further expands the horizon of large language models and empowers exciting developments that involve diverse data beyond text.

The purpose of this paper is to provide a comprehensive survey of the existing research on ChatGPT and its potential applications in various

* Corresponding author.
E-mail address: bob_ge@snnu.edu.cn (B. Ge).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.metrad.2023.100017
Received 7 June 2023; Received in revised form 10 August 2023; Accepted 10 August 2023
Available online 18 August 2023

fields. To achieve this goal, we conducted a thorough analysis of papers related to ChatGPT in the arXiv repository. As of April 1st, 2023, there are a total of 194 papers mentioning ChatGPT on arXiv. In this study, we conducted a trend analysis of these papers and generated a word cloud to visualize the commonly used terms. Additionally, we also examined the distribution of the papers across various fields and presented the corresponding statistics. Fig. 1 displays the submission trend of papers related to ChatGPT, indicating a growing interest in this field. Fig. 2 illustrates the word cloud analysis of all the papers. We can observe that the current research is primarily focused on natural language processing, but there is still significant potential for research in other fields such as education, medical and history. This is further supported by Fig. 3, which displays the distribution of submitted papers across various fields, highlighting the need for more research and development in these areas. Due to the rapid advancement in research related to ChatGPT, we have also introduced a dynamic webpage that provides real-time updates on the latest trends in these area. Interested readers can access the webpage and stay informed about the evolving research directions by following this link.[2]

This paper aims to shed light on the promising capabilities of ChatGPT and provide insight into its potential impact in the future, including ethical considerations. Through this survey, we hope to provide insights into how these models can be improved and extended in the future. In section 2, we will review the existing work related to ChatGPT, including its applications and ethical considerations. In section 3, we conducted a review of existing literature that assesses the capabilities of ChatGPT. We comprehensively evaluated the performance of ChatGPT based on these studies. In addition to discussing the current state of research related to ChatGPT, we will also explore its limitations in section 4. Furthermore, we will provide guidance on future directions for language model development.

2. Related work of ChatGPT

In this section, we review the latest research related to the application and ethics of ChatGPT. Fig. 4 shows the overall framework of this part.

2.1. Application of ChatGPT

2.1.1. Question and answering

2.1.1.1. In the field of education. ChatGPT is commonly used for question and answers testing in the education sector. Users can use ChatGPT to learn, compare and verify answers for different academic subjects such as physics, mathematics, and chemistry, and/or conceptual subjects such as philosophy and religion. Additionally, users can ask open-ended and analytical questions to understand the capabilities of ChatGPT.

In the field of mathematics, Frieder et al.[8] constructed the GHOSTS natural language dataset, which consists of graduate-level math test questions. The authors tested ChatGPT's math abilities on the GHOSTS dataset using a question-and-answer format and evaluated it according to fine-grained standards.In the Grad Text dataset, which covers simple set theory and logic problems, ChatGPT performed the best. However, in the Olympiad-Problem-Solving dataset, ChatGPT performed poorly, receiving only two 4-point scores (out of a total of 5), with the majority of scores being 2 points. In the Holes-in-Proofs dataset, ChatGPT received the lowest score of 1 point. In the MATH dataset, ChatGPT only scored impressively in 26% of cases. These results suggest that ChatGPT's math abilities are clearly lower than those of ordinary math graduate students. Although ChatGPT can generally understand math problems, it fails to provide the correct solutions. Pardos et al.[9] used the Open Adaptive Tutoring system (OATutor) to investigate whether prompts generated by ChatGPT were helpful for learning algebra, with 77 participants from Mechanical Turk taking part in the experiment. The experiment used questions from OpenStax's Elementary and Intermediate Algebra textbooks. These participants were randomly assigned to either a control group (with manual prompts) or an experimental group (with ChatGPT prompts). For each question in both courses, the authors obtained answers from ChatGPT through a question-and-answer format and evaluated scores according to three criteria: ChatGPT provided an answer, the answer was correct, and inappropriate language was not used in the answer. The study found that 70% of prompts generated by ChatGPT passed manual quality checks, and both humans and ChatGPT produced positive learning gains. However, the scores of human prompts ranged from 74.59% to 84.32%, significantly higher than those of ChatGPT prompts. Shakarian et al.[10] studied the performance of ChatGPT on math word problems (MWPs), using the DRAW-1K dataset for experimentation. The dataset consists of 1000 MWPs and their answers, along with algebraic equation templates for solving such problems. The authors used the idea of machine learning introspection and built performance prediction models using random forests and XGBoost, and evaluated them on the dataset using five-fold cross-validation. ChatGPT's accuracy increased from an initial 34% to a final 69%, while its recall increased from an initial 41% to a final 83%. The authors also found that ChatGPT's failure rate decreased from an initial 84% to a final 20%, indicating that performance can vary greatly depending on specific job requirements.

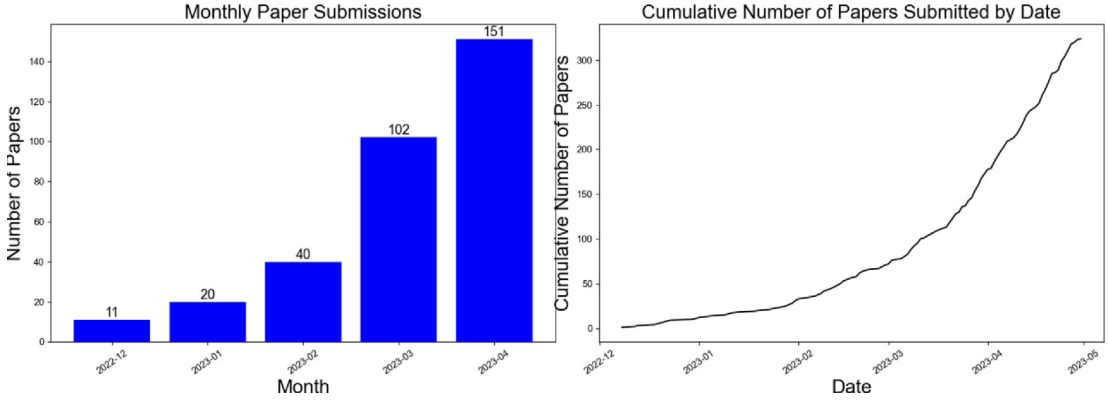

Fig. 1. The graphical representation is utilized to depict the number of research articles related to ChatGPT published from 2022 to April 2023, revealing the trend and growth of ChatGPT-related research over time. The graph showcases the monthly count of submissions and cumulative daily submitted count in arXiv. Over time, there has been an increasing amount of research related to ChatGPT.

[2] https://snnubiai.github.io/chatgpt_arxiv_analysis/.

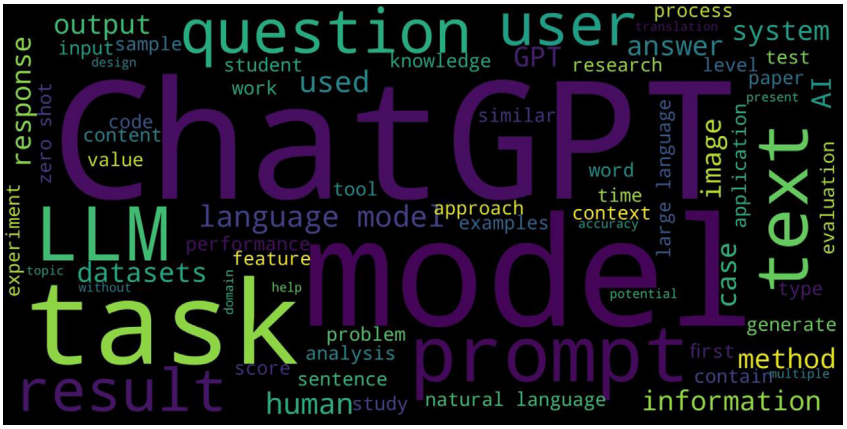

Fig. 2. Word cloud analysis of all the 194 papers.

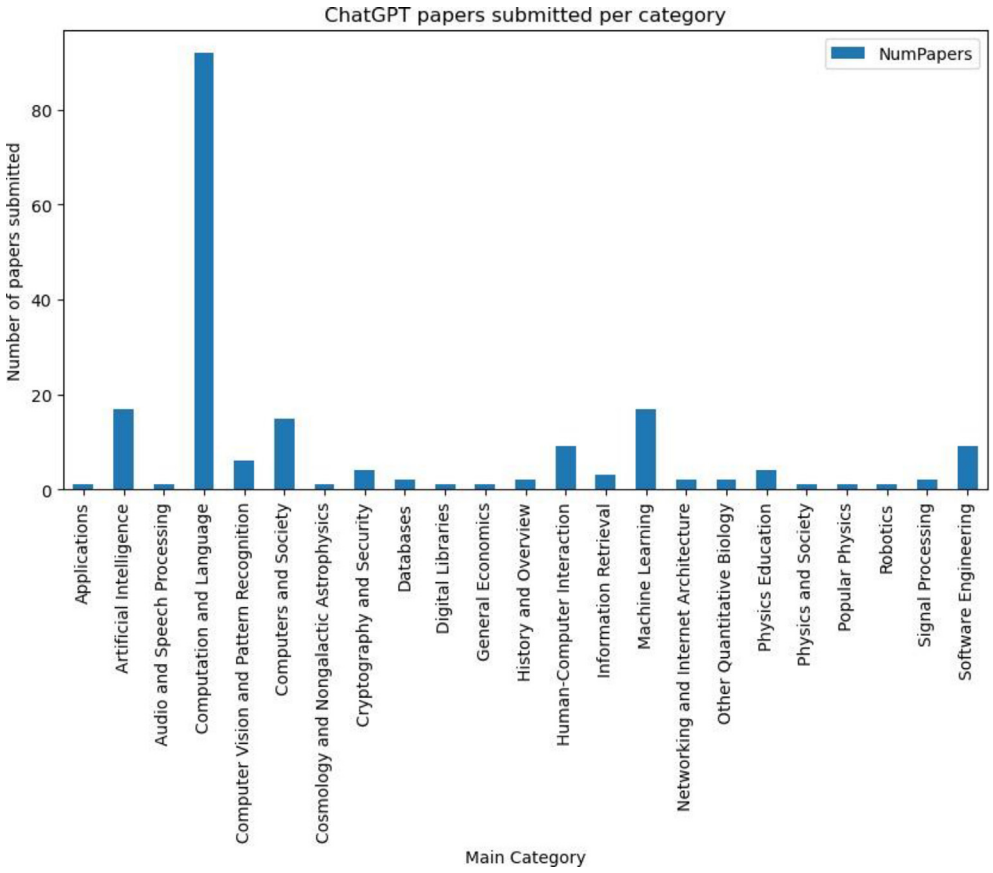

Fig. 3. The distribution of ChatGPT papers submitted across various fields.

In the field of physics, Lehnert et al.[11] explored the capabilities and limitations of ChatGPT by studying how it handles obscure physics topics such as the swamp land conjecture in string theory. The experimental dialogue began with broader and more general questions in the field of string theory before narrowing down to specific swamp land conjectures and examining ChatGPT's understanding of them. The study found that ChatGPT could define and explain different concepts in various styles, but was not effective in truly connecting various concepts. It would confidently provide false information and fabricate statements when necessary, indicating that ChatGPT cannot truly create new knowledge or establish new connections. However, in terms of identifying analogies and describing abstract concepts of visual representation, ChatGPT can cleverly use language. Kortemeyer et al.[12] evaluated ChatGPT's ability to answer calculus-based physics questions through a question-and-answer test. The tests included online homework, clicker questions, programming exercises, and exams covering classical mechanics, thermodynamics, electricity and magnetism, and modern physics. While ChatGPT was able to pass the course, it also demonstrated many misconceptions and errors commonly held by beginners. West et al.[13] used the Force Concept Inventory (FCI) to evaluate ChatGPT's accuracy in answering physics concept problems related to kinematics and Newtonian mechanics in the first semester of college physics. The FCI covers topics such as kinematics, projectile motion, free fall, circular motion, and Newton's laws. The study included data from 415 students who took the FCI at the end of the semester, with an average score of 56%, while ChatGPT scored approximately between 50% to 65%. The authors demonstrated that ChatGPT's performance in physics

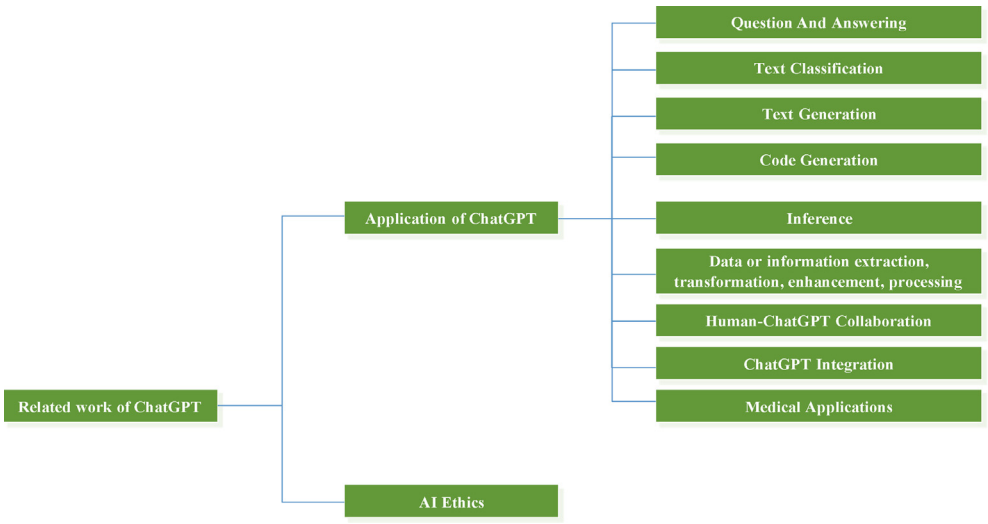

Fig. 4. Structure Diagram of Chapter 2.

learning can reach or even exceed the average level of a semester of college physics.

2.1.1.2. In the medical field. ChatGPT's question-answering capabilities can also be applied in the medical field, such as for answering medical questions from patients or assisting healthcare professionals in diagnosing diseases. Nov et al.[14] evaluated the feasibility of using ChatGPT for patient-doctor communication. The experiment extracted 10 representative patient-doctor interactions from EHR, placed the patient's questions in ChatGPT, and asked ChatGPT to respond using roughly the same number of words as the doctor's response. Each patient's question was answered by either the doctor or ChatGPT, and the patient was informed that 5 were answered by the doctor and 5 were generated by ChatGPT, and was asked to correctly identify the source of the response. The results of the experiment showed that the probability of correctly identifying ChatGPT's response was 65.5%, while the probability of correctly identifying the doctor's response was 65.1%. In addition, the experiment found that the patient's response to the trustworthiness of ChatGPT's function was weakly positive (average Likert score: 3.4), and trust decreased as the complexity of health-related tasks in the questions increased. ChatGPT's responses to patient questions were only slightly different from those of doctors, but people seem to trust ChatGPT to answer low-risk health questions, while for complex medical questions, people still tend to trust the doctor's responses and advice.

Tu et al.[15] explored the causal discovery ability of ChatGPT in the diagnosis of neuropathic pain. Causal relationship discovery aims to reveal potential unknown causal relationships based purely on observed data.[16] The experimental results found that ChatGPT has some limitations in understanding new knowledge and concepts beyond the existing textual training data corpus, that is, it only understands language commonly used to describe situations and not underlying knowledge. In addition, its performance consistency and stability are not high, as the experiment observed that it would provide different answers for the same question under multiple inquiries. However, despite the many limitations of ChatGPT, we believe that it has a great opportunity to improve causal relationship research.

2.1.1.3. In other fields. Guo et al.[17] attempted to apply ChatGPT in the field of communication, specifically using ChatGPT for ordered importance semantic communication, where ChatGPT plays the role of an intelligent consulting assistant that can replace humans in identifying the semantic importance of words in messages and can be directly embedded into the current communication system. For a message to be transmitted, the sender first utilizes ChatGPT to output the semantic importance order of each word. Then, the transmitter executes an unequal error protection transmission strategy based on the importance order to make the transmission of important words in the message more reliable. The experimental results show that the error rate and semantic loss of important words measured in the communication system embedded with ChatGPT are much lower than those of existing communication schemes, indicating that ChatGPT can protect important words well and make semantic communication more reliable.

Wang et al.[18] studied the effectiveness of ChatGPT in generating high-quality Boolean queries for systematic literature search. They designed a wide range of prompts and investigated these tasks on more than 100 systematic review topics. In the end, queries generated by ChatGPT achieved higher accuracy compared to the currently most advanced query generation methods but at the cost of reduced recall. For time-limited rapid reviews, it is often acceptable to trade off higher precision for lower recall. Additionally, ChatGPT can generate high search accuracy Boolean queries by guiding the prompts. However, it should be noted that when two queries use the same prompts, ChatGPT generates different queries, indicating its limitations in consistency and stability. Overall, this study demonstrated the potential of ChatGPT in generating effective Boolean queries for systematic literature searches.

2.1.2. Text classification

The purpose of text classification is to assign text data to predefined categories. This task is critical for many applications, including sentiment analysis, spam detection, and topic modeling. While traditional machine learning algorithms have been widely used for text classification, recent advances in natural language processing have led to the development of more advanced techniques. ChatGPT has shown immense potential in this field. Its ability to accurately classify text, flexibility in handling various classification tasks, and potential for customization make it a valuable tool for text classification, as evidenced by several studies in the literature.

Kuzman et al.[19] employed ChatGPT for automatic genre recognition, with the goal of simplifying the text classification task by utilizing ChatGPT's zero-shot classification capability. They compared ChatGPT's genre recognition performance, using two prompt languages (EN and SL), with the X-GENRE classifier based on the multilingual model XLM-RoBERTa on the English dataset EN-GINCO and the Slovenian dataset GINCO. The results showed that when EN was used as the prompt language, ChatGPT achieved Micro F1, Macro F1, and Accuracy scores of 0.74, 0.66, and 0.72. However, on the GINCO dataset, ChatGPT's genre recognition performance with both EN and SL prompt languages was lower than that of the X-GENRE classifier to varying degrees.

Amin et al.[20] evaluated the text classification ability of ChatGPT in affective computing by using it to perform personality prediction, sentiment analysis, and suicide ideation detection tasks. They prompted ChatGPT with corresponding prompts on three datasets: First Impressions, Sentiment140, and Suicide and Depression, and compared its classification performance with three baseline models: RoBERTa-base, Word2Vec, and BoW. The results showed that ChatGPT's accuracy and UAR for the five personality classifications on the First Impressions dataset were lower than the baseline methods to varying degrees. On the Sentiment140 dataset, ChatGPT's accuracy and UAR were 85.5 and 85.5, respectively, which were better than the three baseline methods. On the Suicide and Depression dataset, ChatGPT's accuracy and UAR were 92.7 and 91.2, respectively, which were lower than RoBERTa, the best-performing baseline method.

Zhang et al.[21] employed ChatGPT for stance detection, which includes support and opposition. They used ChatGPT to classify the political stance of tweets in the SemEval-2016 and P-Stance datasets. SemEval-2016 contains 4870 English tweets, and they selected tweets with the most commonly occurring FM, LA, and HC political labels for stance classification. The P-Stance dataset has 21,574 English tweets, and they classified the stance of tweets towards Trump, Biden, and Bernie. The final results showed that on the SemEval-2016 dataset, ChatGPT achieved F1-m scores of 68.4, 58.2, and 79.5 for the FM, LA, and HC political labels, and F1-avg scores of 72.6, 59.3, and 78.0, respectively. On the P-Stance dataset, ChatGPT achieved F1-m scores of 82.8, 82.3, and 79.4 for the Trump, Biden, and Bernie political figures, and F1-avg scores of 83.2, 82.0, and 79.4, respectively.

Huang et al.[22] used ChatGPT to detect implicit hate speech in tweets. They selected 12.5% (795 tweets) of the LatentHatred dataset containing implicit hate speech and asked ChatGPT to classify them into three categories: implicit hate speech, non-hate speech, and uncertain. The results showed that ChatGPT correctly recognized 636 (80%) of the tweets. The number of tweets classified as non-hate speech and uncertain were 146 (18.4%) and 13 (1.6%), respectively. The results of the reclassification of tweets in the non-hate speech and uncertain categories by Amazon Mechanical Turk (Mturk) workers were consistent with ChatGPT's classification.

Overall, ChatGPT has tremendous potential in text classification tasks, as it can effectively address problems such as genre identification, sentiment analysis, stance detection, and more. However, there are still challenges that ChatGPT faces in the field of text classification. Firstly, it struggles to perform well in classification tasks with rare or out-of- vocabulary words since it heavily relies on the distribution of training data. Additionally, the significant computational resources required for training and utilizing ChatGPT can limit its use in some applications.

2.1.3. Text generation

We live in an era of information explosion, and text is an efficient way of transmitting information. The diversity of information has led to a diversity of text categories. When researchers use ChatGPT's text generation capabilities for research, they inevitably choose to generate different types of text. In the process of reading papers, we found that the word count of the text generated by researchers increased from small to large, so we wanted to summarize existing research based on the size of the text word count. We divided the generated text into three levels: phrases, sentences, and paragraphs.

The following article uses ChatGPT to generate phrases. Zhang et al.[23] proves that the semantic HAR model with semantic augmentation added during training performs better in motion recognition than other models. Semantic augmentation requires shared tokens, which is lacking in some datasets. Therefore, authors leverage ChatGPT for an automated label generation approach for datasets originally without shared tokens. Fu et al.[24] described a new workflow for converting natural language commands into Bash commands. The author uses ChatGPT to generate a candidate list of Bash commands based on user input, and then uses a combination of heuristic and machine learning techniques to rank and select the most likely candidates. This workflow was evaluated on a real command dataset and achieved high accuracy compared to other state-of-the-art methods. Chen et al.[25] used the Bart model and ChatGPT for the task of summarizing humorous titles and compared the performance of the two models. It was found that the Bart model performed better on large datasets, but ChatGPT was competitive with our best fine-tuned model in a small range (48), albeit slightly weaker.

The following article uses ChatGPT to generate sentences.Chen et al.[26] constructed a dialogue dataset (HPD) with scenes, timelines, character attributes, and character relationships in order to use ChatGPT as a conversational agent to generate dialogue. However, ChatGPT's performance on the test set was poor, and there is room for improvement. In study,[27] chatGPT demonstrated its ability to simplify complex text by providing three fictional radiology reports to chatGPT for simplification. Most radiologists found the simplified reports to be accurate and complete, with no potential harm to patients. However, some errors, omissions of critical medical information and text passages were identified, which could potentially lead to harmful conclusions if not understood by the physicians. Xia et al.[28] proposed a new program repair paradigm called Session-based Automated Program Repair (APR). In APR, the previously generated patches are iteratively built upon by combining them with validation feedback to construct the model's input. The effectiveness of the approach is verified using the QuixBugs dataset. The experiment shows that ChatGPT fine-tuned with reinforcement learning from human feedback (RLHF) outperforms Codex trained unsupervisedly in both repair datasets. In reference to study,[29] ChatGPT was compared to three commercial translation products: Google Translate2, DeepL Translate3, and Tencent TranSmart4. The evaluation was conducted on the Flores101 test set, using the WMT19 biomedical translation task to test translation robustness, with BLEU score as the main metric. The study found that ChatGPT is competitive with commercial translation products on high-resource European languages but falls behind on low-resource or distant languages. The authors explored an interesting strategy called pivot prompts, which significantly improved translation performance. While ChatGPT did not perform as well as commercial systems on biomedical abstracts or Reddit comments, it may be a good speech translator. Prieto et al.[30] evaluated the use of ChatGPT in developing an automated construction schedule based on natural language prompts. The experiment required building new partitions in an existing space and providing details on the rooms to be partitioned. The results showed that ChatGPT was able to generate a coherent schedule that followed a logical approach to meet the requirements of the given scope. However, there were still several major flaws that would limit the use of this tool in real-world projects.Michail et al.[31] proposed a method to improve the prediction accuracy of the HeFit fine-tuned XLM_T model on tweet intimacy by generating a dataset of tweets with intimacy rating tags using ChatGPT. The specific operation is to input tweets with intimacy rating tags into ChatGPT and then output similar tweets.

The following article uses ChatGPT to generate paragraphs. Wang et al.[32] compared the abstract summarization performance of ChatGPT and other models on various cross-lingual text datasets and found that ChatGPT may perform worse in metrics such as R_1, R_2, R_L, and B_S. Yang et al.[33] summarized the performance of ChatGPT in question answering-based text summarization and found that, compared to fine-tuned models, ChatGPT's performance is slightly worse in all performance metrics. However, the article suggests that if the dataset is golden annotation, ChatGPT's performance may surpass fine-tuned models in these metrics. Belouadi et al.[34] compared the ability of ByGPT5 and ChatGPT trained on a range of labeled and unlabeled datasets of English and German poetry to generate constrained style poetry, and evaluated them using three metrics: Rhyme, Score-Alliteration, and ScoreMeter Score. The conclusion is that ByGPT5 performs better than ChatGPT. Blanco-Gonzalez et al.[35] evaluated chatGPT's ability to write commentary articles, and in fact, this article itself was written by chatGPT. The human author rewrote the manuscript based on chatGPT's draft. Experts found that it can quickly generate and optimize

text, as well as help users complete multiple tasks. However, in terms of generating new content, it is not ideal. Ultimately, it can be said that without strong human intervention, chatGPT is not a useful tool for writing reliable scientific texts. It lacks the knowledge and expertise required to accurately and fully convey complex scientific concepts and information. Khalil et al.[36] on the originality of content generated by ChatGPT. To evaluate the originality of 50 papers on various topics generated by ChatGPT, two popular plagiarism detection tools, Turnitin and iThenticate, were used. The results showed that ChatGPT has great potential in generating complex text output that is not easily captured by plagiarism detection software. The existing plagiarism detection software should update their plagiarism detection engines. Basic et al.[37] conducted a comparison of the writing performance of students using or not using ChatGPT-3 as a writing aid. The experiment consisted of two groups of 9 participants each. The control group wrote articles using traditional methods, while the experimental group used ChatGPT as an aid. Two teachers evaluated the papers. The study showed that the assistance of ChatGPT did not necessarily improve the quality of the students' essays.Noever et al.[38] discusses the potential of using artificial intelligence (AI), particularly language models like GPT (including GPT-3), to create more convincing chatbots that can deceive humans into thinking they are interacting with another person. The article describes a series of experiments in which they used GPT-3 to generate chatbot responses that mimic human-like conversations and were tested on human participants. The results show that some participants were unable to distinguish between the chatbot and a real human, highlighting the potential for these AI chatbots to be used for deceptive purposes.

2.1.4. Code generation

Code generation refers to the process of automatically generating computer code from high-level descriptions or specifications. ChatGPT's advanced natural language processing capabilities make it capable of performing code generation tasks. By analyzing the requirements for code generation, ChatGPT can produce code snippets that accurately execute the intended functionality. This not only saves time and effort in writing code from scratch but also reduces the risk of errors that may occur during manual coding. In addition, ChatGPT's ability to learn and adapt to new programming languages and frameworks enables it to complete more complex programming tasks. For example: Megahed et al.[39] discussed the potential of using ChatGPT for tasks such as code explanation, suggesting alternative methods for problem-solving with code, and translating code between programming languages. The solutions provided by ChatGPT were found to be viable. In another study, Treude et al.[40] introduced a ChatGPT-based prototype called GPTCOMCARE, which helps programmers generate multiple solutions for a programming problem and highlight the differences between each solution using colors. Sobania et al.[41] utilized ChatGPT for code bug fixing, and further improved the success rate of bug fixing by inputting more information through its dialogue system. Specifically, the QuixBugs standard bug fixing benchmark contained 40 code bugs that needed to be fixed. With limited information, ChatGPT fixed 19 bugs, which was slightly lower than the 21 bugs fixed by the Codex model, but significantly higher than the 7 fixed by the Standard APR model. When given more prompts and information, ChatGPT was able to fix 31 bugs, demonstrating its potential for code bug fixing tasks. Xia et al.[28] proposed a conversational approach for Automate Program Repair (APR), which alternates between generating patches and validating them against feedback from test cases until the correct patch is generated. Selecting 30 bugs from the QuixBugs standard bug fixing benchmark, which are suitable for test case feedback, and demonstrating them with Java and Python, the QuixBugs-Python and QuixBugs-Java datasets were obtained. The conversational APR using ChatGPT outperformed the conversational APR using Codex and the conversational APR using CODEGEN (with model parameters of 350 M, 2 B, 6 B, and 16 B) on both datasets. Furthermore, ChatGPT's conversational APR generated and validated patches with significantly fewer feedback loops than the other models.

ChatGPT can not only be used to achieve some simple code generation tasks but also can be used to accomplish some complex programming tasks. Noever et al.[42] tested ChatGPT's code generation capabilities on four datasets - Iris, Titanic, Boston Housing, and Faker. When prompted to mimic a Python interpreter in the form of a Jupyter notebook, the model was able to generate independent code based on the prompt and respond with the expected output. For example, when given the prompt "data.cor ()" for the Iris dataset, ChatGPT generated correct Python output. The test results indicate that ChatGPT can access structured datasets and perform basic software operations required by databases, such as create, read, update, and delete (CRUD). This suggests that cutting-edge language models like ChatGPT have the necessary scale to tackle complex problems. McKee et al.[43] utilized ChatGPT as an experimental platform to investigate cybersecurity issues. They modeled five different modes of computer virus properties, including self-replication, self-modification, execution, evasion, and application, using ChatGPT. These five modes encompassed thirteen encoding tasks from credential access to defense evasion within the MITRE ATT&CK framework. The results showed that the quality of ChatGPT's generated code was generally above average, except for the self-replication mode, where it performed poorly.They[44] also employed ChatGPT as a network honeypot to defend against attackers. By having ChatGPT mimic Linux, Mac, and Windows terminal commands and providing interfaces for TeamViewer, nmap, and ping, a dynamic environment can be created to adapt to attackers' operations, and logs can be used to gain insight into their attack methods, tactics, and procedures. The authors demonstrated ten honeypot tasks to illustrate that ChatGPT's interface not only provides sufficient API memory to execute previous commands without defaulting to repetitive introductory tasks but also offers a responsive welcome program that maintains attackers' interest in multiple queries.

In the field of code generation, there are still several challenges with ChatGPT. Firstly, its application scope is limited as its training data is biased towards programming languages such as Python, C++, and Java, making it potentially unsuitable for some programming languages or coding styles. Secondly, manual optimization is necessary for code formatting, as the generated code may not be performance-optimized or follow best coding practices, requiring manual editing and optimization. Lastly, the quality of the generated code cannot be guaranteed, as it heavily relies on the quality of the natural language input, which may contain errors, ambiguities, or inconsistencies, ultimately affecting the accuracy and reliability of the generated code.

2.1.5. Inference

Inference refers to the process of drawing new conclusions or information through logical deduction from known facts or information. It is typically based on a series of premises or assumptions, and involves applying logical rules or reasoning methods to arrive at a conclusion. Inference is an important ability in human thinking, and is often used to solve problems, make decisions, analyze and evaluate information, etc. Inference also plays a key role in fields such as science, philosophy, law, etc. There are two types of inference: inductive reasoning, which involves deriving general rules or conclusions from known facts or experiences, and deductive reasoning, which involves deriving specific conclusions from known premises or assumptions. Whether inductive or deductive, the process of inference requires following strict logical rules to ensure the correctness and reliability of the inference.

Some papers attempt to use ChatGPT's ability in inductive reasoning to capture the meaning in text and use defined metrics to score the text. Michail et al.[31] uses ChatGPT to infer intimacy expressed in tweets. They first input 50 tweets with intimacy markers to ChatGPT, then use inductive reasoning to infer the standards for generating tweets with different levels of intimacy, and finally generate ten tweets with intimacy values ranging from 0 to 5. Susnjak et al.[45] collected a large amount of

textual data from patient-doctor discussion forums, patient testimonials, social media platforms, medical journals, and other scientific research publications. Using the BERT model, the author inferred emotion values from 0 to 1. The author visualized the process of how the presence of bias in the discourse surrounding chronic manifestations of the disease using the SHAP tool. The author also envisioned ChatGPT as a replacement for the BERT model for scoring the emotional value of text. Huang et al.[22] chose 12.5% of individuals in the potential hate dataset as study materials, induced ChatGPT to make classifications based on a prompt, and ChatGPT produced three classifications: unclear, yes, and no. The author assigned a value of 1 to yes, −1 to no, and 0 to unclear, and had ChatGPT score and classify them. ChatGPT was able to correctly classify 80% of implicit hate tweets in the author's experimental setup, demonstrating ChatGPT's great potential as a data labeling tool using simple prompts.

Some papers have evaluated ChatGPT's reasoning performance, mainly in decision-making and spatial reasoning, and identifying ambiguity. Tang et al.[46] used the independence axiom and the transitivity axiom, as well as other non-VNM related decision-making abilities, by presenting bets conditioned on random events, bets with asymmetric outcomes, decisions encapsulating Savage's Sure Thing principle, and other complex bet structures like nested bets, to design experiments where each experiment input a short prompt to ChatGPT and evaluated the results. The conclusion is that ChatGPT exhibits uncertainty in the decision-making process: in some cases, large language models can arrive at the correct answer through incorrect reasoning; and it may make suboptimal decisions for simple reasoning problems. Ortega-Martn et al.[47] had ChatGPT detect three different levels of language ambiguity and evaluated its performance. The conclusion is that In semantics, ChatGPT performed perfectly in the detection of ambiguities. Apart from that, it has some bright sports (co-reference resolution) and some weaknesses (puts gender bias over grammar in some non-ambiguous situations). In the generation task ChatGPT did well, but also revealed some of its worse issues: the lack of systematicity. Lastly, it should also be pointed that in most of the cases ChatGPT brilliantly alludes to lack of context as the key factor in disambiguation.

2.1.6. Data or information extraction, transformation, enhancement, processing

2.1.6.1. Data visualization. Natural language interfaces have contributed to generating visualizations directly from natural language, but visualization problems remain challenging due to the ambiguity of natural language.ChatGPT provides a new avenue for the field by converting natural language into visualized code.

In terms of data visualization, Noever et al.[42] tested ChatGPT's basic arithmetic skills by asking questions.On the iris dataset, Titanic survival dataset, Boston housing data, and randomly generated insurance claims dataset, the statistical analysis of data and visualization problems were converted to programming problems using Jupyter to verify ChatGPT's ability to generate python code to draw suitable graphs and analyze the data. The results show that ChatGPT can access structured and organized datasets to perform the four basic software operations required for databases: create, read, update, and delete, and generate suitable python code to plot graphs for descriptive statistics, variable correlation analysis, describing trends, and other data analysis operations.Maddigan et al.[48] proposed an end-to-end solution for visualizing data in natural language using LLM, which uses an open-source python framework designed to generate appropriate hints for selected datasets to make LLM more effective in understanding natural language, and uses internal reasoning capabilities to select the appropriate visualization type to generate the code for visualization. In this paper, the reseachers compare the visualization results of GPT-3, Codex and ChatGPT in the case of nvBench SQLite database[49] and the visualization results of energy production dataset in the study of ADVISor with NL4DV.[50,51]In addition to, they explore the ability to reason and hypothesize of the LLM on movie dataset[49] when the hints are insufficient or wrong.Experimental results show that LLM can effectively support the end-to-end generation of visualization results from natural language when supported by hints, providing an efficient, reliable and accurate solution to the natural language visualization problem.

2.1.6.2. Information extraction. The goal of information extraction is to extract specific information from natural language text for structured representation, including three important subtasks such as entity relationship extraction, named entity recognition, and event extraction, which have wide applications in business, medical, and other fields.

In information extraction, Wei et al.[52] proposed ChatIE, a ChatGPT-based multi-round question-and-answer framework for information extraction. The framework decomposes a complex information extraction (IE) task into several parts, then combines the results of each round into a final structured result. The entity association triple extraction, named entity recognition, and event extraction tasks were performed on six datasets NYT11-HRL, DuIE2.0, conllpp, MSR, DuEE1.0,[53–55,56,57] and ACE05 in both languages, comparing three metrics of precision, recall, and F1 score.These results suggest that on six widely used IE datasets, ChatIE improves performance by an average of 18.98% compared to the original ChatGPT without ChatIE, and outperforms the supervised models FCM and MultiR[58,59] on the NYT11-HRL dataset.While the original ChatGPT cannot solve complex IE problems with original task instructions, and with this framework, successfully IE tasks were implemented on six datasets.Gao et al.[60] explored the feasibility and challenges of ChatGPT for event extraction on the ACE2005 corpus, evaluating the performance of ChatGPT in long-tail and complex scenarios (texts containing multiple events) and comparing it with two task-specific models, Text2Event and EEQA.[61,62]Then, they explored the impact of different cues on performance of ChatGPT. The results show that the average performance of ChatGPT in long-tail and complex scenarios is only 51.04% of that of task-specific models such as EEQA. Continuous refinement of cues does not lead to consistent performance improvements, and ChatGPT is highly sensitive to different cue styles. Tang et al.[63] proposed a new training paradigm that incorporates appropriate cues to guide ChatGPT to generate a variety of examples with different sentence structures and language patterns and eliminate the resulting low-quality or duplicate samples for downstream tasks. Although compared to a soft model for a specific healthcare task, ChatGPT underperforms in Named Entity Recognition (NER) and Relationship Extraction (RE) tasks, in the Gene Association Database (GAD) Release; EU-ADR corpus for the RE task, the innovative training framework was able to train local models, with F1 scores improving from 23.37% to 63.99% for the named entity recognition task and from 75%, while alleviating privacy concerns and time-consuming data collection and annotation problems.He et al.[64] proposed a contextual learning framework ICL- D3IE. this framework introduces formatted presentation, continuously iterates to update and improve the presentation, and then combines ChatGPT for text information extraction. In the paper, ICL-D3IE is compared with existing pre-trained models such as LiLT, BROS (in-distribution (ID) setting and out-of-distribution (OOD) setting) on datasets (FUNSD, CORD, and SROIE[65,66,67]).These results show that the ICL-D3IE method in all datasets and settings except for the ID setting on CORD are superior to other methods, with ICL-D3IE (GPT-3) F1 scores reaching 90.32% on FUNSD and97.88% on SROIE; in the out-of-distribution (OOD) setting, ICL-D3IE performs much better than previous pre-trained methods on all datasets.Polak et al.[68] proposed ChatExtract method - consisting of a set of engineering prompts applied to a conversational LLM - for automatic data extraction. During experiment, they extracted a large number of sentences from hundreds of papers and randomly selected 100 sentences containing data and 100 sentences without data as test data. The results show that the accuracy and recall of LLM exceeded 90% and may be comparable to human accuracy in many cases; in addition to this, the experiments were conducted under the condition of removing follow-up prompts and not keeping the conversation compared to previous experiments, respectively.

The accuracy of deleting follow-up questions dropped to 80.2% and the recall rate dropped to 88.0%. Removing the conversational aspect and related information retention recall and accuracy dropped to 90.0% and 56.6%, respectively, demonstrating the effect of information retention combined with purposeful redundancy on LLM information extraction performance.

2.1.6.3. Quality assessment. For translation quality, text generation quality, manual assessment is usually effective but suffers from subjectivity and time-consuming, etc. It was found through exploration that ChatGPT has also achieved significant performance in automatic quality assessment.

In terms of quality assessment, Kocmi et al.[69] proposed a GPT-based translation quality assessment metric, GEMBA, which evaluates the translation of each fragment individually and then averages all the obtained scores to obtain a final system-level score. In the MQM2022 test set (English-German, English-Russian, and Chinese-English),[70] a scoring task was performed with a classification task to compare the accuracy[71] and kendall tau scores[72] of seven GPT models under four cue templates.The results showed that GEMBA had the highest system-level accuracy of 88.0% compared to more than 10 automatic metrics such as BLEU, and among the seven GPT models, ChatGPT accuracy is above 80%, in addition to, the best performance can be obtained in the least constrained template, demonstrating the potential of LLM for translation quality assessment tasks, but the evaluation is only applicable at the system level and needs further improvement.Wang et al.[73] used ChatGPT as a natural language generation (NLG) evaluator to study the correlation with human judgment. On three datasets covering different NLG tasks, task- and aspect-specific cues were designed to guide ChatGPT for NLG evaluation in CNN/DM,[74] OpenMEVA- ROC, and BAGEL for summary, story generation, and data-to-text scoring, respectively. Then, they compute Spearman coefficients,[75] Pearson correlation coefficients.[76] Kendall's Tau score[77] to assess the correlation with human evaluations.The results show that ChatGPT is highly correlated with human judgments in all aspects, with correlation coefficients of 0.4 or more in all categories, showing its potential as an NLG indicator.

2.1.6.4. Data augmentation. In natural language processing, text data augmentation is an effective measure to alleviate the problem of low data quantity and low quality training data, and ChatGPT has shown great potential in this regard.

In terms of data augmentation, Dai et al.[78] proposed a ChatGPT-based text data augmentation method that reformulates each sentence in the training sample into multiple conceptually similar but semantically different samples for classification tasks downstream of the Bert model.On text transcriptions and PubMed 20k datasets containing more than 8 h of audio data of common medical symptom descriptions, experiments were conducted to compare cosine similarity and TransRate metrics with multiple data enhancement methods.[79] This paper shows that compared with existing data enhancement methods, the proposed ChatAug method shows a double-digit improvement in sentence classification accuracy and generates more diverse augmented samples while maintaining its accuracy, but the original model is not fine-tuned in the paper and suffers from a lack of domain knowledge, which may produce incorrect augmented data.

2.1.6.5. Multimodal fusion. ChatGPT can currently only process natural language directly, but with a cross-modal encoder, it can combine natural language with cross-modal processing to provide solutions for intelligent transportation, healthcare, and other fields.

In terms of multimodal data processing, Wu et al.[80] constructed a framework that Visual ChatGPT integrates with different Visual Foundation Models (VFMs) and then combines a series of hints to input visual information to ChatGPT to solve visual problems.The paper shows examples of visual tasks such as removing or replacing certain objects from images, interconversion between images and text, demonstrating the Visual ChatGPT has great potential and capability for different tasks.But there are issues during the task that requires a large number of hints to convert VFMs to language, invoke multiple VFMs to solve complex problems leading to limited real-time capability, and security and privacy issues. Zheng et al.[81] showed a text mining example of LLM for extracting self-driving car crash data from California crash news, analyzing a failure report example, and generating a crash report example based on keywords; introduced a use case concept of a smartphone-based framework for automatic LLM failure report generation, which absorbs multiple data sources captured by cell phone sensors and then transfers the data to a language space for text mining, inference and generation, and further outputs the key information needed to form a comprehensive fault report, demonstrating the potential of LLM for a variety of transportation tasks.

Nowadays, ChatGPT shows a wide range of applications in data visualization, information extraction, data enhancement, quality assessment, and multimodal data processing.But there are also issues on how to further utilize hints to effectively interact with ChatGPT, lack of ability to process and analyze data from devices such as sensors, and data privacy and security.

2.1.6.6. Cueing techniques. Cue engineering provides important support for effective dialogue with large language models.White et al.[82] proposed a framework for cueing models applicable to different domains. This framework structures cues to interact with LLMs by providing specific rules and guidelines. Also, this paper presents a catalog of cueing patterns that have been applied to LLM interactions, as well as specific examples with and without cues. The advantages of the combinability of prompting patterns are demonstrated, allowing users to interact with LLM more effectively, but patterns for reusable solutions and new ways to use LLM need to be continuously explored.

2.1.7. Human-ChatGPT collaboration

Collaboration between humans and machines is a process where humans and machines work together to achieve a common goal. In such collaboration, humans provide domain expertise, creativity, and decision-making abilities, while machines provide automation, scalability, and computing power. ChatGPT is an advanced natural language processing model that can understand and generate human-like language, thereby reducing communication costs. Its ability to process and generate natural language makes it an ideal partner for human collaboration. ChatGPT can offer relevant suggestions, complete tasks based on human input, and enhance human productivity and creativity. It can learn from human feedback and adapt to new tasks and domains, further improving its performance in human-machine collaboration. ChatGPT's capability to comprehend natural language and produce appropriate responses makes it a valuable tool for various collaboration applications, as demonstrated by several studies in the literature we have gathered.

Ahmad et al.[83] proposed a method for human-machine collaboration using ChatGPT to create software archi-tecture. This method transforms software stories (created by software architects based on application scenarios) into feasible software architecture diagrams through continuous interaction between the software architect and ChatGPT. During the evaluation stage, ChatGPT uses the Software Architecture Analysis Method (SAAM) to evaluate each component in the software architecture and generate evaluation reports. This method efficiently utilizes the knowledge and supervision of the architect with the capabilities of ChatGPT to collaboratively build software-intensive systems and services. Lanzi et al.[84] proposed a collaborative design framework that combines interactive evolution and ChatGPT to simulate typical human design processes. Humans collaborate with large language models (such as ChatGPT) to recombine and transform ideas, and use genetic algorithms to iterate through complex creative tasks.

The results of three game design tasks showed that the framework received positive feedback from game designers. The framework has

good reusability and can be applied to any design task that can be described in free text form.

In the future, ChatGPT's ability to understand nonverbal cues such as tone of voice and body language can be enhanced, enabling it to better understand human thoughts and interact with people more effectively.

2.1.8. ChatGPT integration

Integration refers to combining different systems or software components to achieve a common goal. ChatGPT can be integrated as a part of a whole or act as an integration tool to enable seamless communication between different systems. Its natural language processing ability makes it easier for non-technical users to interact with systems, reducing the need for specialized knowledge or training. Some studies in the literature we collected have already demonstrated this.

Treude et al.[40] integrated ChatGPT into the prototype of "GPTCOMCARE" to address programming query problems. This integration allowed for the generation of multiple source code solutions for the same query, which increased the efficiency of software development. The results of their study demonstrated the effectiveness of using ChatGPT to improve the quality and diversity of code solutions, ultimately reducing the amount of time and effort required for software development. Wang et al.[85] proposed the chatCAD method, which utilizes large language models (LLMs) such as ChatGPT to enhance the output of multiple CAD networks for medical images, including diagnosis, lesion segmentation, and report generation networks. The method generates suggestions in the form of a chat dialogue. The authors tested the effectiveness of the method on a randomly selected set of 300 cases from the MIMIC-CXR dataset, which included 50 cases each of cardiomegaly, edema, consolidation, atelectasis, pleural effusion, and no findings. Compared to CvT2DistilGPT2 and R2GenCMN, chatCAD showed significant advantages in RC and F1, while only performing weaker than R2GenCMN in PR.

Integrating ChatGPT into applications will still present challenges. Firstly, ChatGPT's performance may be affected by language barriers or differences in terminology between different systems. Additionally, ChatGPT's responses are not always deterministic, which poses a challenge when integrating with systems that require precise and reproducible results. Finally, the processing time of ChatGPT is slow for integration tasks involving time-sensitive data such as traffic, which is a limitation in time-critical environments.

2.1.9. Medical applications

ChatGPT offers promising applications in medical field, revolutionizing healthcare practices. Its natural language processing capabilities enable interactive assistance for radiologists, aiding in image annotation, lesion detection, and classification. ChatGPT's extensive knowledge base facilitates real-time feedback, context-specific recommendations, and streamlined report generation. By integrating ChatGPT into workflows, healthcare professionals benefit from enhanced efficiency and precision in clinical decision-making, fostering accessible and collaborative healthcare solutions. For example:

ChatCAD[85] integrates large language models (LLMs) into computer-aided diagnosis (CAD) networks for medical imaging. It has shown promising results in improving diagnosis, lesion segmentation, and report generation, three key aspects of CAD networks. This integration represents a notable effort in combining large language models with medical imaging techniques.

Hu et al.[86] conducted a comprehensive review of language models in the context of medical imaging and highlighted the potential advantages of ChatGPT in enhancing clinical workflow efficiency, reducing diagnostic errors, and supporting healthcare professionals. Their work aims to bridge the gap between large language models and medical imaging, paving the way for new ideas and innovations in this research domain.

Ma et al.[87] proposed ImpressionGPT, a novel approach that harnesses the powerful in-context learning capabilities of ChatGPT. They achieve this by creating dynamic contexts using domain-specific and individualized data. The dynamic prompt method enables the model to learn contextual knowledge from semantically similar examples in existing data and iteratively optimize the results, aiding radiologists in composing the "impression" section based on the "findings" section. The results demonstrate state-of-the-art performance on both the MIMIC-CXR and OpenI datasets, without the need for additional training data or fine-tuning of the LLMs.

AD-AutoGPT,[88] an integration of AutoGPT,[89] leverages the power of ChatGPT in an automated processing pipeline that can assist users in accomplishing nearly any given task. With AD-AutoGPT, users can autonomously generate data collection, processing, and analysis pipelines based on their text prompts. Through AD-AutoGPT, detailed trend analysis, mapping of topic distances, and identification of significant terms related to Alzheimer's disease (AD) have been achieved from four new sources specifically relevant to AD. This significantly contributes to the existing knowledge base and facilitates a nuanced understanding of discourse surrounding diseases in the field of public health. It lays the groundwork for future research in AI-assisted public health studies.

Patient privacy protection has always been a significant concern in the healthcare field. DeID-GPT[90] aims to explore the potential of ChatGPT in the de-identification and anonymization of medical reports. Experimental results demonstrate that ChatGPT exhibit promising capabilities in medical data de-identification compared to other LLMs.

Despite notable efforts, the integration of large language models and medical imaging still presents several challenges. Firstly, the intricate and technical nature of medical imaging data, which encompasses detailed anatomical structures and subtle abnormalities, may not be effectively conveyed or comprehended through the text-based chat interface of large language models. Secondly, ChatGPT lacks the specialized medical knowledge and training necessary for precise interpretation and analysis of medical images, potentially leading to dangerous misunderstandings or inaccurate diagnoses.[91] It is imperative to establish various machine learning models to detect samples generated by both humans and ChatGPT, in order to prevent false medical information produced by ChatGPT from causing misjudgments in disease progression, delaying treatment processes, or negatively impacting patients' lives and health. Lastly, the legal and ethical aspects associated with deploying artificial intelligence models like ChatGPT in a medical context, such as patient privacy and liability concerns, must be thoughtfully addressed and aligned with regulatory standards. While ChatGPT is powerful, it is not easily applicable in clinical settings. Compliance with HIPAA regulations, privacy issues, and the necessity for IRB approval pose significant obstacles,[90] primarily because these models require uploading patient data to external hosting platforms. One possible solution to this problem is to address it through localized deployment of language models, such as Radiology-GPT.[92] The future application of chatGPT in the field of medical imaging will necessitate ongoing efforts from all stakeholders.

2.2. AI ethics

Since the advent of ChatGPT, this powerful natural language processing model has not only brought great convenience to people but also triggered more crisis-aware thinking. Some researchers have started to hypothesize and study the potential negative impacts of ChatGPT. This proactive research provides good proposals for standardized construction to address future AI abuse issues.

Regarding the possibility of ChatGPT being used for plagiarism and cheating, Zhou et al.[93] reflected on the current state of development of artificial intelligence like ChatGPT. As ChatGPT becomes increasingly easy to obtain and scalable in text generation, there is a high likelihood that these technologies will be used for plagiarism, including scientific literature and news sources, posing a great threat to the credibility of various forms of news media and academic articles. Some scholars are concerned that the end of paper as a meaningful evaluation tool may be approaching,[94,95] as ChatGPT can easily generate persuasive paragraphs,

chapters, and papers on any given topic. Additionally, it will exacerbate plagiarism issues in many fields such as education, medicine, and law,[11] and may be used for cheating in academic exams.[96] Definitional recognition technology is a relatively effective method for detecting plagiarism, and the definitional typology proposed in Ref. [93] can alleviate people's concerns by being used to construct new datasets. Susnjak[96] proposed a solution to the possibility of large language models like ChatGPT being used for exam cheating: guiding ChatGPT to generate some critical thinking problems through questioning, then providing answers and critically evaluating them. Analysis of ChatGPT shows that it exhibits critical thinking, can generate highly realistic text in terms of accuracy, relevance, depth, breadth, logic, persuasiveness, and originality. Therefore, educators must be aware of the possibility of ChatGPT being used for exam cheating and take measures to combat cheating behavior to ensure the fairness of online exams.

Regarding the evaluation of ChatGPT's own political and ethical tendencies, Hartmann et al.[97] used Wahl-*O*- Mat, one of the most commonly used voting advice applications in the world, to show ChatGPT political statements from different parties, forcing it to make choices of agree, disagree, or neutral. The results indicated that ChatGPT has a pro-environment, left-wing liberal ideology, which was also confirmed in the nation-state agnostic political compass test. Another study (referenced as[98]) examined ChatGPT's moral standards by repeatedly asking it different versions of the trolley problem, and found that ChatGPT gave answers with different moral orientations, lacking a firm moral stance. A subsequent test also found that ChatGPT's lack of consistency could affect people's moral judgments. Additionally, Borji et al.[99] demonstrated ChatGPT's inconsistency in reasoning, factual errors, mathematics, coding, and bias across eleven related aspects. These findings highlight ChatGPT's inherent traits and limitations, and people should be aware of their potential impact when seeking advice from ChatGPT. Zhuo et al.[100] comprehensively analyzed the moral hazard, bias, reliability, robustness, and toxicity of ChatGPT from four perspectives. The results found that ChatGPT may perform slightly better than the current SOTA language model, but has some shortcomings in all four aspects. The authors look ahead to the ethical challenges of developing advanced language models and suggest directions and strategies for designing ethical language models.

Regarding relevant policies and regulations, Hacker et al.[101] discussed the nature and rules of large generative AI models, including ChatGPT, which are rapidly changing the way we communicate, explain, and create. The author suggested that different stakeholders in the value chain should take regulatory responsibility and deploy four strategies to tailor more comprehensive laws for the benefit of society. Another study (referenced as[102]) criticized the European Commission's proposal on AI responsibility and suggested revising the proposed AI responsibility framework to ensure effective compensation while promoting innovation, legal certainty, and sustainable AI regulation. A policy framework was proposed (referenced as[103]) to customize LLMs, such as ChatGPT, in a socially acceptable and safe manner, emphasizing the need to align large language models (LLMs) with human preferences.

The political and ethical tendencies of ChatGPT could influence users' behavior and decision-making to some extent. However, some studies have conducted in-depth research on the use of norms and limitations, which could enable humans to use ChatGPT more reasonably and safely.

3. Evaluation

3.1. Comparison of ChatGPT with existing popular models

We use publicly available datasets to comprehensively evaluate the strengths and limitations of ChatGPT. Reference [104] evaluates the technical performance of ChatGPT in multitask, multilingual, and multimodal aspects based on 23 standard public datasets and newly designed multimodal datasets, including eight different common natural language processing application tasks. The experimental results show that, in terms of multitasking, ChatGPT outperforms various state-of-the-art zero-shot learning large language models in most tasks, and even outperforms fine-tuned task-specific models in some individual tasks. In terms of multilingualism, we found that ChatGPT cannot be applied to low-resource languages because it cannot understand the language and generate translations for that language. In terms of multimodality, ChatGPT's ability is still basic compared to specialized language-visual models.

In terms of stability, reference [105] concludes that ChatGPT's performance is always lower than SOTA, the current state-of-the-art model, in almost all tasks. This means that as a general model, ChatGPT has never reached the level of the best existing models. Experimental data shows that the average quality of the SOTA model is 73.7%, while the average quality of the ChatGPT model is only 56.5%. At the same time, ChatGPT's stability is poor: the standard deviation of its performance is 23.3%, while the SOTA model's standard deviation is only 16.7%. This non-deterministic behavior exhibited by ChatGPT could be a serious drawback in some problems.

Similarly, Qin et al.[106] conducted a comprehensive evaluation of whether ChatGPT is a qualified general natural language processing task solver. The experiment analyzed ChatGPT's zero-shot learning ability based on 20 commonly used public datasets covering 7 representative task categories. Below, we will analyze ChatGPT's performance on each task:

In terms of reasoning tasks, ChatGPT performs average on mathematical symbol, commonsense causal, and logical reasoning tasks, but performs well in arithmetic reasoning.[106] That is to say, ChatGPT's abilities vary among different types of reasoning tasks. In terms of logical reasoning, ChatGPT's deductive and abductive reasoning are superior to inductive reasoning, while in other reasoning tasks, such as analogy, causal and commonsense reasoning, ChatGPT performs well.[104]

In terms of sentiment analysis task, ChatGPT performs similarly to GPT-3.5 and bert-style models.[106,107] However, according to literature,[105] ChatGPT has losses not exceeding 25% on most tasks, except for three relatively subjective emotion perception tasks where it performs poorly. If we remove these tasks to calculate the average quality of the two models, we find that the SOTA method has an average quality of 80%, while the ChatGPT method has an average quality of 69.7%. That is to say, ChatGPT performs well on all tasks except for emotion-related tasks, and can handle most of the problems we consider. However, overall, its performance is lower than the SOTA model based on experimental data, but the difference between the two is not very large.

In other tasks, according to literature,[106] ChatGPT performs well in natural language inference, i.e., the task of inferring sentence relationships, and its performance on this task is significantly better than all bert-style models.[107] However, while ChatGPT performs well on inference tasks, it may produce some self-contradictory or unreasonable responses, which is its potential limitation. In question-answering, dialogue, and summarization tasks, ChatGPT performs better than the GPT-3.5 model,[106] especially in the question-answering task, where its performance is comparable to bert-style models.[107] Therefore, we have demonstrated that ChatGPT is a qualified general-purpose model.

However, ChatGPT also has limitations in many aspects. Firstly, it lacks the ability to handle non-textual semantic reasoning tasks such as mathematical, temporal, and spatial reasoning, and it performs poorly in multi-hop reasoning.[104] Secondly, ChatGPT is not good at solving named entity recognition tasks.[106] Furthermore, ChatGPT performs poorly in handling tasks involving negative connotations and neutral similarity.[107] Finally, these conclusions indicate that, like other large pre-trained language models, ChatGPT has limitations in completing complex reasoning tasks.

In summary, ChatGPT's zero-shot performance is comparable to fine-tuned bert and GPT-3.5 models, and with the help of advanced prompting strategies, ChatGPT can demonstrate better comprehension abilities. However, it still cannot outperform the current SOTA models.

3.2. Feedback from ChatGPT users

In response to feedback from ChatGPT users, Haque et al.[108] conducted a mixed-methods study using 10,732 early ChatGPT user tweets. The authors extracted Twitter data using Python and Twitter API and constructed the ChatGPTTweet dataset, which contains 18k tweets. For each tweet, the authors collected information on text content, user location, occupation, verification status, date of publication, and tags. Based on this dataset, the authors studied the characteristics of early ChatGPT users, discussion topics related to ChatGPT on Twitter, and the sentiment of Twitter users toward ChatGPT. For RQ1, the authors found that early ChatGPT users had a diverse and wide range of occupational backgrounds and geographical locations. For RQ2, the authors identified nine topics related to ChatGPT, including its impact on software development, entertainment and creativity, natural language processing, education, chatbot intelligence, business development, search engines, question-answering tests, and future careers and opportunities. For RQ3, most early users expressed positive sentiment toward topics such as software development and creativity, while only a few expressed concern about the potential misuse of ChatGPT.

3.3. Adverse effects of ChatGPT on users

Regarding the negative effects of ChatGPT on users, Luan et al.[109] studied the psychological principles of ChatGPT, delved into the factors that attract users' attention, and revealed the impact of these factors on future learning. In the post-pandemic era, teachers and students are both facing uncertainty in the teaching process and job pressures. Under these common constraints of education and employment, educators and students must re-evaluate current educational methods and outcomes, as well as students' future career development. Through question-and-answer exchanges with ChatGPT, people can easily obtain appropriate solutions or key information, thereby enhancing their motivation, eliminating anxiety in learning, improving interest, and achieving psychological satisfaction. Subhash et al.[110] explored whether large language models have the ability to reverse user preferences. With the development of pre-trained large language models, people are increasingly concerned about the ability of these models to influence, persuade, and potentially manipulate user preferences in extreme cases. Therefore, the literature[110] roughly qualitatively analyzed that adversarial behavior does lead to potential changes in user preferences and behaviors in dialogue systems. If we want to further quantitatively analyze the ability of large language models in this regard, additional statistical summary techniques need to be used for future research.

4. Discussion

4.1. Limitations

Despite the remarkable capabilities of ChatGPT, it still faces certain limitations. Some of these limitations include.

4.1.1. Outdated knowledge

The current models are trained on historical data (up to 2021), thereby lacking real-time comprehension of current affairs. This is a critical concern in today's information-explosion era, as the reliability of prior knowledge bases progressively diminishes, potentially yielding inaccurate responses, especially in rapidly evolving domains such as jurisprudence and technology. Additionally, these models are incapable of fact-checking while the training data is composed of content from various sources, some of which may be unreliable, which may result in seemingly plausible yet nonsensical responses.

4.1.2. Insufficient understanding

While these models can interpret the majority of inquiries and contextual situations, they occasionally encounter comprehension biases when addressing ambiguous or contextually complex queries. Furthermore, in certain spe-cialized fields, the abundance of unique abbreviation exacerbates the models' understanding challenges, resulting in incorrect and vacuous responses.

4.1.3. Energy consumption

Throughout the training and inference stages, these large-scale models require significant computational resources and electrical power, resulting in elevated energy consumption and significant carbon emissions. Consequently, this restricts their deployment and practical applications.

4.1.4. Malicious usage

Despite OpenAI implementing a series of restrictions to mitigate model toxicity, instances of users evading these constraints through meticulously designed prompts have emerged, inducing the model to produce unhealthy content or even using it for illicit commercial purposes.

4.1.5. Bias and discrimination

Due to the influence of pre-training data, the models exhibit biases in political, ideological, and other areas. The application of LLMs in public domains, such as education and publicity, should be approached with extreme caution.

4.1.6. Privacy and data security

Concurrent with the expansion of users, protecting user privacy and data security becomes increasingly important. In fact, ChatGPT was banned in Italy in early April due to privacy concerns. This is particularly relevant given the models' extensive collection of personal information and preferences during interactions, and as future multimodal models, such as GPT-4, may frequently require users to upload private photos.

4.2. Future directions

In forthcoming research, the development of models based on ChatGPT may focus on addressing these limitations to enhance their practical applications.

Primarily, researchers should continue to work on refining model training methodologies while filtering pre-training data to minimize the presence of misleading information in the model's knowledge base, thereby obtaining accurate responses. Concurrently, it is crucial to emphasize training approaches that economize computational resources, thereby mitigating costs and broadening potential application scenarios.

Moreover, the advancements in context-awareness and disambiguation technologies are anticipated to facilitate enhanced comprehension of complex queries by models, improving the accuracy, relevance, and context-awareness of AI-generated content. Integrating real-time data streams can also keep these models in sync with current events and trends, enabling them to provide up-to-date information such as live traffic, weather, and stock updates.

Additionally, developers should engage in interdisciplinary collaboration with specialists from diverse domains, including policy-making, jurisprudence, and sociology, with the objective of formulating standard and ethical frameworks for LLM development, deployment, and utilization, thereby alleviating potential harmful consequences. In terms of public awareness and education, mandatory awareness training should be implemented prior to large-scale public deployment and application to increase public awareness of LLM capabilities and limitations while promoting responsible and informed utilization, especially in industries such as K-12 education and journalism.

Furthermore, ChatGPT still lacks specific domain knowledge and may encounter potential data security issues, especially in the medical field. In domains where error tolerance is low and data privacy and security are crucial, such as medical applications,[90] localized training and deployment of LLMs should be considered.[92] Customizing training for specific LLMs based on domain-specific data should also be taken into account.

Finally, the influence of ChatGPT should not be limited to just the NLP field. They also show promising prospects in the areas of computer vision, brain-inspired AI, and robotics. These models exhibit a capacity for learning and comprehension comparable with human-level intelligence, positioning them as a pivotal component in the development of artificial general intelligence (AGI).[111] Their ability to facilitate seamless interactions between humans and robots paves the way for the execution of more complex tasks. The remarkable capacity of zero-shot in-context learning of these models enables quick adaptation to new tasks without the requirement for labeled data for fine-tuning, which is a critical challenge in fields like medical informatics[90] and robotics[112] where the availability of labeled data is commonly limited or non-existent.

5. Conclusion

This review paper provides a comprehensive survey of ChatGPT, highlighting their potential applications and significant contributions to the field of natural language processing. The findings of this study reveal that the interest in these models is growing rapidly, and they have shown considerable potential for application across a wide range of domains. One key factor contributing to the success of ChatGPT is their ability to perform large-scale pre-training, which captures knowledge from the vast expanse of the internet, allowing the models to learn from a massive amount of data. The integration of Reinforcement Learning from Human Feedback (RLHF) has further enhanced the model's adaptability and performance, making it highly efficient in processing natural language. In addition, RLHF aligns language models with human preferences & values and empower text generation with the naturalness of human style. This study has also identified several potential ethical concerns related to the development and use of ChatGPT. For instance, there are concerns about the generation of biased or harmful content, privacy violations, and the potential for misuse of the technology. It is crucial to address these concerns and ensure that ChatGPT is developed and used in a responsible and ethical manner. Furthermore, the results of this study demonstrate that there is significant potential for ChatGPT to be applied in a range of domains, including education, medical, history, mathematics, physics, and more. These models can facilitate tasks such as generating summaries, answering questions, and providing personalized recommendations to users. Overall, the insights presented in this review paper can serve as a useful guide for researchers and practitioners looking to advance the field of natural language processing. Future research in this field should focus on addressing ethical concerns, exploring new applications, and ensuring the responsible use of ChatGPT. The potential of these models to revolutionize natural language processing is enormous, and we look forward to seeing more developments in this field.

Authorship statement

All authors wrote the manuscript and have approved the final version of the manuscript.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61976131).

References

1. Brown Tom, Mann Benjamin, Ryder Nick, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
2. Ouyang Long, Wu Jeff, Jiang Xu, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. 2022.
3. Radford Alec, Narasimhan Karthik, Tim Salimans, et al. *Improving Language Understanding by Generative Pre-training*. OpenAI; 2018.
4. Radford Alec, Wu Jeffrey, Amodei Dario, et al. Better language models and their implications. *OpenAI Blog*. 2019;1(2). https://openai.com/blog/better-language-models.
5. Radford Alec, Wu Jeffrey, Child Rewon, et al. Language models are unsupervised multitask learners. *OpenAI blog*. 2019;1(8):9.
6. OpenAI. *Gpt-4 Technical Report*. 2023.
7. Christiano Paul F, Jan Leike, Brown Tom, Martic Miljan, Shane Legg, Amodei Dario. Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst*. 2017;30.
8. Frieder Simon, Pinchetti Luca, Griffiths Ryan-Rhys, et al. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*. 2023.
9. Pardos Zachary A, Bhandari Shreya. Learning gain differences between chatgpt and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*. 2023.
10. Shakarian Paulo, Koyyalamudi Abhinav, Noel Ngu, Mareedu Lakshmivihari. An independent evaluation of chatgpt on mathematical word problems (mwp). *arXiv preprint arXiv:2302.13814*. 2023.
11. Kay Lehnert. Ai insights into theoretical physics and the swampland program: a journey through the cosmos with chatgpt. *arXiv preprint arXiv:2301.08155*. 2023.
12. Kortemeyer Gerd. Could an artificial-intelligence agent pass an introductory physics course? *arXiv preprint arXiv:2301.12127*. 2023.
13. West Colin G. Ai and the fci: can chatgpt project an understanding of introductory physics? *arXiv preprint arXiv:2303.01067*. 2023.
14. Nov Oded, Singh Nina, Devin M, Mann. *Putting Chatgpt's Medical Advice to the (Turing) Test*. medRxiv; 2023.
15. Tu Ruibo, Ma Chao, Zhang Cheng. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis. *arXiv preprint arXiv:2301.13819*. 2023.
16. Clark Glymour, Zhang Kun, Spirtes Peter. Review of causal discovery methods based on graphical models. *Front Genet*. 2019;10:524.
17. Guo Shuaishuai, Wang Yanhu, Li Shujing, Saeed Nasir. Semantic communications with ordered importance using chatgpt. *arXiv preprint arXiv:2302.07142*. 2023.
18. Wang Shuai, Scells Harrisen, Koopman Bevan, Guido Zuccon. *Can Chatgpt Write a Good Boolean Query for Systematic Review Literature Search?*. 2023. *arXiv preprint arXiv:2302.03495*.
19. Kuzman Taja, Mozetic Igor, Ljubešic Nikola. Chatgpt: beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *arXiv e-prints*. 2023:2303.
20. Amin Mostafa M, Cambria Erik, W Schuller Björn. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *arXiv preprint arXiv:2303.03186*. 2023.
21. Zhang Bowen, Ding Daijun, Jing Liwen. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*. 2022.
22. Huang Fan, Kwak Haewoon, An Jisun. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*. 2023.
23. Zhang Xiyuan, Chowdhury Ranak Roy, Hong Dezhi, Gupta Rajesh K, Shang Jingbo. Modeling label semantics improves activity recognition. *arXiv preprint arXiv:2301.03462*. 2023.
24. Fu Quchen, Teng Zhongwei, Georgaklis Marco, White Jules, C Schmidt Douglas. Nl2cmd: an updated workflow for natural language to bash commands translation. *arXiv preprint arXiv:2302.07845*. 2023.
25. Chen Yanran, Eger Steffen. Transformers go for the lols: generating (humourous) titles from scientific abstracts end-to-end. *arXiv preprint arXiv:2212.10522*. 2022.
26. Chen Nuo, Wang Yan, Jiang Haiyun, Deng Cai, Chen Ziyang, Jia Li. What would harry say? building dialogue agents for characters in a story. *arXiv preprint arXiv:2211.06869*. 2022.
27. Jeblick Katharina, Schachtner Balthasar, Dexl Jakob, Andreas Mittermeier, Anna Theresa Stüber, Topalis Johanna, Weber Tobias, Wesp Philipp, Sabel Bastian, Ricke Jens, et al. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*. 2022.
28. Chunqiu Steven Xia, Zhang Lingming. Conversational automated program repair. *arXiv preprint arXiv:2301.13246*. 2023.
29. Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Zhaopeng Tu. Is Chatgpt a Good Translator? Yes with Gpt-4 as the Engine.
30. Prieto Samuel A, Mengiste Eyob T, García de Soto Borja. Investigating the use of ChatGPT for the scheduling of construction projects. *Buildings*. mar 2023;13(4):857.
31. Michail Andrianos, Konstantinou Stefanos, Simon Clematide. Uzh_clyp at semeval-2023 task 9: head-first fine-tuning and chatgpt data generation for cross-lingual learning in tweet intimacy prediction. *arXiv preprint arXiv:2303.01194*. 2023.
32. Wang Jiaan, Liang Yunlong, Meng Fandong, Li Zhixu, Qu Jianfeng, Zhou Jie. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*. 2023.
33. Yang Xianjun, Li Yan, Zhang Xinlu, Chen Haifeng, Cheng Wei. *Exploring the Limits of Chatgpt for Query or Aspect-Based Text Summarization*. 2023. *arXiv preprint arXiv:2302.08081*.
34. Jonas Belouadi, Eger Steffen. Bygpt5: end-to-end style-conditioned poetry generation with token-free language models. *arXiv preprint arXiv:2212.10474*. 2022.
35. Blanco-Gonzalez Alexandre, Cabezon Alfonso, Seco-Gonzalez Alejandro, et al. The role of ai in drug discovery: challenges, opportunities, and strategies. *arXiv preprint arXiv:2212.08104*. 2022.
36. Khalil Mohammad, Er Erkan. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*. 2023.

37. Basic Zeljana, Banovac Ana, Kruzic Ivana, Jerkovic Ivan. Better by you, better than me, chatgpt3 as writing assistance in students essays. *arXiv preprint arXiv:2302.04536*. 2023.
38. Noever David, Matt Ciolino. The turing deception. *arXiv preprint arXiv:2212.06721*. 2022.
39. Megahed Fadel M, Chen Ying-Ju, Ferris Joshua A, Knoth Sven, Jones-Farmer L Allison. How generative ai models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study. *arXiv preprint arXiv:2302.10916*. 2023.
40. Treude Christoph. *Navigating Complexity in Software Engineering: A Prototype for Comparing Gpt-N Solutions*. 2023. *arXiv preprint arXiv:2301.12169*.
41. Sobania Dominik, Briesch Martin, Hanna Carol, Petke Justyna. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*. 2023.
42. Noever David, McKee Forrest. Numeracy from literacy: data science as an emergent skill from large language models. *arXiv preprint arXiv:2301.13382*. 2023.
43. McKee Forrest, Noever David. Chatbots in a botnet world. *arXiv preprint arXiv:2212.11126*. 2022.
44. McKee Forrest, Noever David. Chatbots in a honeypot world. *arXiv preprint arXiv:2301.03771*. 2023.
45. Teo Susnjak. Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. *arXiv preprint arXiv:2302.06474*. 2023.
46. Tang Zhisheng, Kejriwal Mayank. *A Pilot Evaluation of Chatgpt and Dall-E 2 on Decision Making and Spatial Reasoning*. 2023. *arXiv preprint arXiv:2302.09068*.
47. Ortega-Martín Miguel, García-Sierra Óscar, Ardoiz Alfonso, Álvarez Jorge, Juan Carlos Armenteros, Alonso Adrián. Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*. 2023.
48. Paula Maddigan, Teo Susnjak. Chat2vis: generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. *arXiv preprint arXiv:2302.02094*. 2023.
49. Luo Yuyu, Tang Jiawei, Li Guoliang. nvbench: a large-scale synthesized dataset for cross-domain natural language to visualization task. *arXiv preprint arXiv:2112.12926*. 2021.
50. Liu Can, Han Yun, Jiang Ruike, Yuan Xiaoru. Advisor: automatic visualization answer for natural-language question on tabular data. In: *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE; 2021:11–20.
51. Narechania Arpit, Srinivasan Arjun, Stasko John. Nl4dv: a toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Trans Visual Comput Graph*. 2020;27(2):369–379.
52. Xiang Wei, Cui Xingyu, Cheng Ning, et al. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*. 2023.
53. Levow Gina-Anne. The third international Chinese language processing bakeoff: word segmentation and named entity recognition. In: *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. 2006:108–117.
54. Li Shuangjie, He Wei, Shi Yabing, et al. Duie: a large-scale Chinese dataset for information extraction. In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II*. vol. 8. Springer; 2019:791–800.
55. Li Xinyu, Li Fayuan, Pan Lu, et al. Duee: a large-scale dataset for Chinese event extraction in real-world scenarios. In: *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II*. vol. 9. Springer; 2020:534–545.
56. Takanobu Ryuichi, Zhang Tianyang, Liu Jiexi, Huang Minlie. A hierarchical framework for relation extraction with reinforcement learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33. 2019:7072–7079.
57. Wang Zihan, Shang Jingbo, Liu Liyuan, Lu Lihao, Liu Jiacheng, Han Jiawei. Crossweigh: training named entity tagger from imperfect annotations. *arXiv preprint arXiv:1909.01441*. 2019.
58. Gormley Matthew R, Yu Mo, Dredze Mark. Improved relation extraction with feature-rich compositional embedding models. *arXiv preprint arXiv:1505.02419*. 2015.
59. Hoffmann Raphael, Zhang Congle, Xiao Ling, Zettlemoyer Luke, Daniel S Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011:541–550.
60. Gao Jun, Zhao Huan, Yu Changlong, Xu Ruifeng. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*. 2023.
61. Du Xinya, Cardie Claire. Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*. 2020.
62. Lu Yaojie, Lin Hongyu, Xu Jin, et al. Text2event: controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*. 2021.
63. Tang Ruixiang, Han Xiaotian, Jiang Xiaoqian, Hu Xia. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*. 2023.
64. He Jiabang, Wang Lei, Hu Yi, et al. Icl-d3ie: in-context learning with diverse demonstrations updating for document information extraction. *arXiv preprint arXiv:2303.05063*. 2023.
65. Huang Zheng, Chen Kai, He Jianhua, et al. Icdar2019 competition on scanned receipt ocr and information extraction. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE; 2019:1516–1520.
66. Jaume Guillaume, Hazim Kemal Ekenel, Thiran Jean-Philippe. Funsd: a dataset for form understanding in noisy scanned documents. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. vol. 2. IEEE; 2019:1–6.
67. Park Seunghyun, Shin Seung, Lee Bado, et al. Cord: a consolidated receipt dataset for post-ocr parsing. In: *Workshop on Document Intelligence at NeurIPS 2019*. 2019.
68. Polak Maciej P, Morgan Dane. Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. *arXiv preprint arXiv:2303.05352*. 2023.
69. Kocmi Tom, Federmann Christian. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*. 2023.
70. Freitag Markus, Rei Ricardo, Mathur Nitika, et al. Results of wmt22 metrics shared task: stop using bleu–neural metrics are better and more robust. In: *Proceedings of the Seventh Conference on Machine Translation*. WMT; 2022:46–68.
71. Kocmi Tom, Federmann Christian, Grundkiewicz Roman, Junczys-Dowmunt Marcin, Matsushita Hitokazu, Menezes Arul. To ship or not to ship: an extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*. 2021.
72. Freitag Markus, Rei Ricardo, Mathur Nitika, et al. Results of wmt22 metrics shared task: stop using bleu–neural metrics are better and more robust. In: *Proceedings of the Seventh Conference on Machine Translation*. WMT; 2022:46–68.
73. Wang Jiaan, Liang Yunlong, Meng Fandong, et al. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*. 2023.
74. Hermann Karl Moritz, Kocisky Tomas, Grefenstette Edward, et al. Teaching machines to read and comprehend. *Adv Neural Inf Process Syst*. 2015;28.
75. H Zar Jerrold. Spearman rank correlation. *Encyclopedia of biostatistics*. 2005;7.
76. Mukaka Mavuto M. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69–71.
77. Kendall Maurice G. A new measure of rank correlation. *Biometrika*. 1938;30(1/2):81–93.
78. Dai Haixing, Liu Zhengliang, Liao Wenxiong, et al. Chataug: leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*. 2023.
79. Huang Long-Kai, Huang Junzhou, Yu Rong, Yang Qiang, Ying Wei. *Frustratingly Easy Transferability Estimation*. 2022:9201–9225.
80. Wu Chenfei, Yin Shengming, Qi Weizhen, Wang Xiaodong, Tang Zecheng, Duan Nan. Visual chatgpt: talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*. 2023.
81. Zheng Ou, Abdel-Aty Mohamed, Wang Dongdong, Wang Zijin, Ding Shengxuan. Chatgpt is on the horizon: could a large language model be all we need for intelligent transportation? *arXiv preprint arXiv:2303.05382*. 2023.
82. White Jules, Fu Quchen, Hays Sam, et al. *A Prompt Pattern Catalog to Enhance Prompt Engineering with Chatgpt*. 2023. *arXiv preprint arXiv:2302.11382*.
83. Ahmad Aakash, Waseem Muhammad, Liang Peng, et al. Towards human-bot collaborative software architecting with chatgpt. *arXiv preprint arXiv*. 2023;2302:14600.
84. Luca Lanzi Pier, Loiacono Daniele. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. *arXiv preprint arXiv:2303.02155*. 2023.
85. Wang Sheng, Zhao Zihao, Ouyang Xi, Wang Qian, Shen Dinggang. Chatcad: interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*. 2023.
86. Hu Mingzhe, Pan Shaoyan, Li Yuheng, Yang Xiaofeng. Advancing medical imaging with language models: a journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*. 2023.
87. Ma Chong, Wu Zihao, Wang Jiaqi, et al. Impressiongpt: an iterative optimizing framework for radiology report summarization with chatgpt. *arXiv preprint arXiv:2304.08448*. 2023.
88. Dai Haixing, Li Yiwei, Liu Zhengliang, et al. Ad-autogpt: an autonomous gpt for alzheimer's disease infodemiology. *arXiv preprint arXiv:2306.10095*. 2023.
89. Gravitas Significant. *Auto-gpt: An Autonomous Gpt-4 Experiment*. 2023.
90. Liu Zhengliang, Yu Xiaowei, Zhang Lu, et al. Deid-gpt: zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*. 2023.
91. Liao Wenxiong, Liu Zhengliang, Dai Haixing, et al. Differentiate chatgpt-generated and human-written medical texts. *arXiv preprint arXiv:2304.11567*. 2023.
92. Liu Zhengliang, Zhong Aoxiao, Li Yiwei, et al. Radiology-gpt: a large language model for radiology. *arXiv preprint arXiv:2306.08666*. 2023.
93. Zhou Chao, Qiu Cheng, Acuna Daniel E. *Paraphrase Identification with Deep Learning: A Review of Datasets and Methods*. 2022. *arXiv preprint arXiv:2212.06933*.
94. de Winter JCF. *Can Chatgpt Pass High School Exams on English Language Comprehension?*. 2023.
95. Yeadon Will, Inyang Oto-Obong, Mizouri Arin, Peach Alex, Craig Testrow. The death of the short-form physics essay in the coming ai revolution. *arXiv preprint arXiv:2212.11661*. 2022.
96. Teo Susnjak. Chatgpt: the end of online exam integrity? *arXiv preprint arXiv:2212.09292*. 2022.
97. Hartmann Jochen, Jasper Schwenzow, Witte Maximilian. The political ideology of conversational ai: converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*. 2023.
98. Krügel Sebastian, Ostermaier Andreas, Uhl Matthias. The moral authority of chatgpt. *arXiv preprint arXiv:2301.07098*. 2023.
99. Ali Borji. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*. 2023.
100. Zhuo Terry Yue, Huang Yujin, Chen Chunyang, Xing Zhenchang. Exploring ai ethics of chatgpt: a diagnostic analysis. *arXiv preprint arXiv:2301.12867*. 2023.
101. Hacker Philipp, Engel Andreas, Mauer Marco. Regulating chatgpt and other large generative ai models. *arXiv preprint arXiv:2302.02337*. 2023.
102. Hacker Philipp. The european ai liability directives–critique of a half-hearted approach and lessons for the future. *arXiv preprint arXiv:2211.13960*. 2022.
103. Kirk Hannah Rose, Vidgen Bertie, Paul Röttger, Hale Scott A. Personalisation within bounds: a risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*. 2023.
104. Bang Yejin, Cahyawijaya Samuel, Lee Nayeon, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*. 2023.
105. Jan Kocoń, Cichecki Igor, Kaszyca Oliwier, et al. Chatgpt: jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*. 2023.

106. Qin Chengwei, Zhang Aston, Zhang Zhuosheng, Chen Jiaao, Yasunaga Michihiro, Yang Diyi. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*. 2023.
107. Zhong Qihuang, Ding Liang, Liu Juhua, Du Bo, Tao Dacheng. *Can Chatgpt Understand Too? a Comparative Study on Chatgpt and Fine-Tuned Bert*. 2023. *arXiv preprint arXiv:2302.10198*.
108. Ul Haque Mubin, Dharmadasa Isuru, Zarrin Tasnim Sworna, Rajapakse Roshan Namal, Ahmad Hussain. i think this is the most disruptive technology": exploring sentiments of chatgpt early adopters using twitter data. *arXiv preprint arXiv: 2212.05856*. 2022.
109. Luan Lingfei, Lin Xi, Li Wenbiao. Exploring the cognitive dynamics of artificial intelligence in the post-covid-19 and learning 3.0 era: a case study of chatgpt. *arXiv preprint arXiv:2302.04818*. 2023.
110. Subhash Varshini. Can large language models change user preference adversarially? *arXiv preprint arXiv:2302.10291*. 2023.
111. Zhao Lin, Zhang Lu, Wu Zihao, et al. When brain-inspired ai meets agi. *arXiv preprint arXiv:2303.15935*. 2023.
112. Liu David, Chen Yuzhong, Zihao Wu. Digital twin (dt)-cyclegan: enabling zero-shot sim-to-real transfer of visual grasping models. *IEEE Rob Autom Lett*. 2023;8(5): 2421–2428.