A Survey on Evaluation of Large Language Models

YUPENG CHANG and XU WANG, School of Artificial Intelligence, Jilin University, Changchun, China
JINDONG WANG, Microsoft Research Asia, Beijing, China
YUAN WU, School of Artificial Intelligence, Jilin University, Changchun, China
LINYI YANG, Westlake University, Hangzhou, Hangzhou, China
KAIJIE ZHU, Institute of Automation, Chinese Academy of Sciences, Beijing, China
HAO CHEN, Carnegie Mellon University, Pittsburgh, USA
XIAOYUAN YI, Microsoft Research Asia, Beijing, China
CUNXIANG WANG, Westlake University, Hangzhou, China
YIDONG WANG and WEI YE, Peking University, Beijing, China
YUE ZHANG, Westlake University, Hangzhou, China
YI CHANG, School of Artificial Intelligence, Jilin University, Changchun, China
PHILIP S. YU, University of Illinois at Chicago, Chicago, USA
QIANG YANG, Hong Kong University of Science and Technology, Kowloon, China
XING XIE, Microsoft Research Asia, Beijing, China

Large language models (LLMs) are gaining increasing popularity in both academia and industry, owing to their unprecedented performance in various applications. As LLMs continue to play a vital role in both research and daily use, their evaluation becomes increasingly critical, not only at the task level, but also at the society level for better understanding of their potential risks. Over the past years, significant efforts have been made to examine LLMs from various perspectives. This paper presents a comprehensive review of these evaluation methods for LLMs, focusing on three key dimensions: *what to evaluate*, *where to evaluate*, and *how to evaluate*. Firstly, we provide an overview from the perspective of evaluation tasks, encompassing general natural language processing tasks, reasoning, medical usage, ethics, education, natural and social sciences, agent applications, and other areas. Secondly, we answer the 'where' and 'how' questions by diving into the evaluation methods and benchmarks, which serve as crucial components in assessing the performance of LLMs. Then, we summarize the success and failure cases of LLMs in different tasks. Finally, we shed light on several future challenges that lie ahead in LLMs evaluation. Our aim is to offer invaluable insights to researchers in the realm of LLMs evaluation, thereby aiding the development of more proficient LLMs. Our key point is that

Y. Chang and X. Wang contributed equally to this research.

This work is supported in part by NSF under grant III-2106758.

Authors' addresses: Y. Chang, X. Wang, Y. Wu (Corresponding author), and Y. Chang, School of Artificial Intelligence, Jilin University, 2699 Qianjin St., Changchun, China, 130012; e-mails: ypchang_jluai@outlook.com, xwang22@mails.jlu.edu.cn, yuanwu@jlu.edu.cn; J. Wang (Corresponding author), X. Yi, and X. Xie, Microsoft Research Asia, Beijing, China; e-mail: Jindong.wang@microsoft.com; L. Yang, C. Wang, and Y. Zhang, Westlake University, Hangzhou, China; K. Zhu, Institute of Automation, Chinese Academy of Sciences, Beijing, China; H. Chen, Carnegie Mellon University, Pennsylvania, USA; Y. Wang and W. Ye, Peking University, Beijing, China; P. S. Yu, University of Illinois at Chicago, Illinois, USA; Q. Yang, Hong Kong University of Science and Technology, Kowloon, Hong Kong, China.

ACM 2157-6904/2024/03-ART39
https://doi.org/10.1145/3641289

evaluation should be treated as an essential discipline to better assist the development of LLMs. We consistently maintain the related open-source materials at: https://github.com/MLGroupJLU/LLM-eval-survey

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning;*

Additional Key Words and Phrases: Large language models, evaluation, model assessment, benchmark

ACM Reference Format:
Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (March 2024), 45 pages. https://doi.org/10.1145/3641289

1 INTRODUCTION

Understanding the essence of intelligence and establishing whether a machine embodies it poses a compelling question for scientists. It is generally agreed upon that authentic intelligence equips us with reasoning capabilities, enables us to test hypotheses, and prepares for future eventualities [92]. In particular, **Artificial Intelligence (AI)** researchers focus on the development of machine-based intelligence, as opposed to biologically based intellect [136]. Proper measurement helps to understand intelligence. For instance, measures for general intelligence in human individuals often encompass IQ tests [12].

Within the scope of AI, the Turing Test [193], a widely recognized test for assessing intelligence by discerning if responses are of human or machine origin, has been a longstanding objective in AI evolution. It is generally believed among researchers that a computing machine that successfully passes the Turing Test can be considered as intelligent. Consequently, when viewed from a wider lens, the chronicle of AI can be depicted as the timeline of creation and evaluation of intelligent models and algorithms. With each emergence of a novel AI model or algorithm, researchers invariably scrutinize its capabilities in real-world scenarios through evaluation using specific and challenging tasks. For instance, the Perceptron algorithm [49], touted as an **Artificial General Intelligence (AGI)** approach in the 1950s, was later revealed as inadequate due to its inability to resolve the XOR problem. The subsequent rise and application of **Support Vector Machines (SVMs)** [28] and deep learning [104] have marked both progress and setbacks in the AI landscape. A significant takeaway from previous attempts is the paramount importance of AI evaluation, which serves as a critical tool to identify current system limitations and inform the design of more powerful models.

Recently, **large language models (LLMs)** have incited substantial interest across both academic and industrial domains [11, 219, 255]. As demonstrated by existing work [15], the great performance of LLMs has raised promise that they could be AGI in this era. LLMs possess the capabilities to solve diverse tasks, contrasting with prior models confined to solving specific tasks. Due to its great performance in handling different applications such as general natural language tasks and domain-specific ones, LLMs are increasingly used by individuals with critical information needs, such as students or patients.

Evaluation is of paramount prominence to the success of LLMs due to several reasons. First, evaluating LLMs helps us better understand the strengths and weakness of LLMs. For instance, the PromptBench [262] benchmark illustrates that current LLMs are sensitive to adversarial prompts, thus a careful prompt engineering is necessary for better performance. Second, better evaluations can provide better guidance for human-LLMs interaction, which could inspire future interaction design and implementation. Third, the broad applicability of LLMs underscores the paramount importance of ensuring their safety and reliability, particularly in safety-sensitive sectors such

as financial institutions and healthcare facilities. Finally, as LLMs are becoming larger with more emergent abilities, existing evaluation protocols may not be enough to evaluate their capabilities and potential risks. Therefore, we aim to raise awareness in the community of the importance to LLMs evaluations by reviewing the current evaluation protocols and most importantly, shed light on future research about designing new LLMs evaluation protocols.

With the introduction of ChatGPT [145] and GPT-4 [146], there have been a number of research efforts aiming at evaluating ChatGPT and other LLMs from different aspects (Figure 2), encompassing a range of factors such as natural language tasks, reasoning, robustness, trustworthiness, medical applications, and ethical considerations. Despite these efforts, a comprehensive overview capturing the entire gamut of evaluations is still lacking. Furthermore, the ongoing evolution of LLMs has also presented novel aspects for evaluation, thereby challenging existing evaluation protocols and reinforcing the need for thorough, multifaceted evaluation techniques. While existing research such as Bubeck et al. [15] claimed that GPT-4 can be seen as sparks of AGI, others contest this claim due to the human-crafted nature of its evaluation approach.

This paper serves as the first comprehensive survey on the evaluation of large language models. As depicted in Figure 1, we explore existing work in three dimensions: 1) What to evaluate, 2) Where to evaluate, and 3) How to evaluate. Specifically, "what to evaluate" encapsulates existing evaluation tasks for LLMs, "where to evaluate" involves selecting appropriate datasets and benchmarks for evaluation, while "how to evaluate" is concerned with the evaluation process given appropriate tasks and datasets. These three dimensions are integral to the evaluation of LLMs. We subsequently discuss potential future challenges in the realm of LLMs evaluation.

The contributions of this paper are as follows:

(1) We provide a comprehensive overview of LLMs evaluations from three aspects: what to evaluate, where to evaluate, and how to evaluate. Our categorization is general and encompasses the entire life cycle of LLMs evaluation.
(2) Regarding what to evaluate, we summarize existing tasks in various areas and obtain insightful conclusions on the success and failure case of LLMs (Section 6), providing experience for future research.
(3) As for where to evaluate, we summarize evaluation metrics, datasets, and benchmarks to provide a profound understanding of current LLMs evaluations. In terms of how to evaluate, we explore current protocols and summarize novel evaluation approaches.
(4) We further discuss future challenges in evaluating LLMs. We open-source and maintain the related materials of LLMs evaluation at https://github.com/MLGroupJLU/LLM-eval-survey to foster a collaborative community for better evaluations.

The paper is organized as follows. In Section 2, we provide the basic information of LLMs and AI model evaluation. Then, Section 3 reviews existing work from the aspects of "what to evaluate". After that, Section 4 is the "where to evaluate" part, which summarizes existing datasets and benchmarks. Section 5 discusses how to perform the evaluation. In Section 6, we summarize the key findings of this paper. We discuss grand future challenges in Section 7 and Section 8 concludes the paper.

2 BACKGROUND

2.1 Large Language Models

Language models (LMs) [36, 51, 96] are computational models that have the capability to understand and generate human language. LMs have the transformative ability to predict the likelihood of word sequences or generate new text based on a given input. N-gram models [13], the most common type of LM, estimate word probabilities based on the preceding context. However, LMs

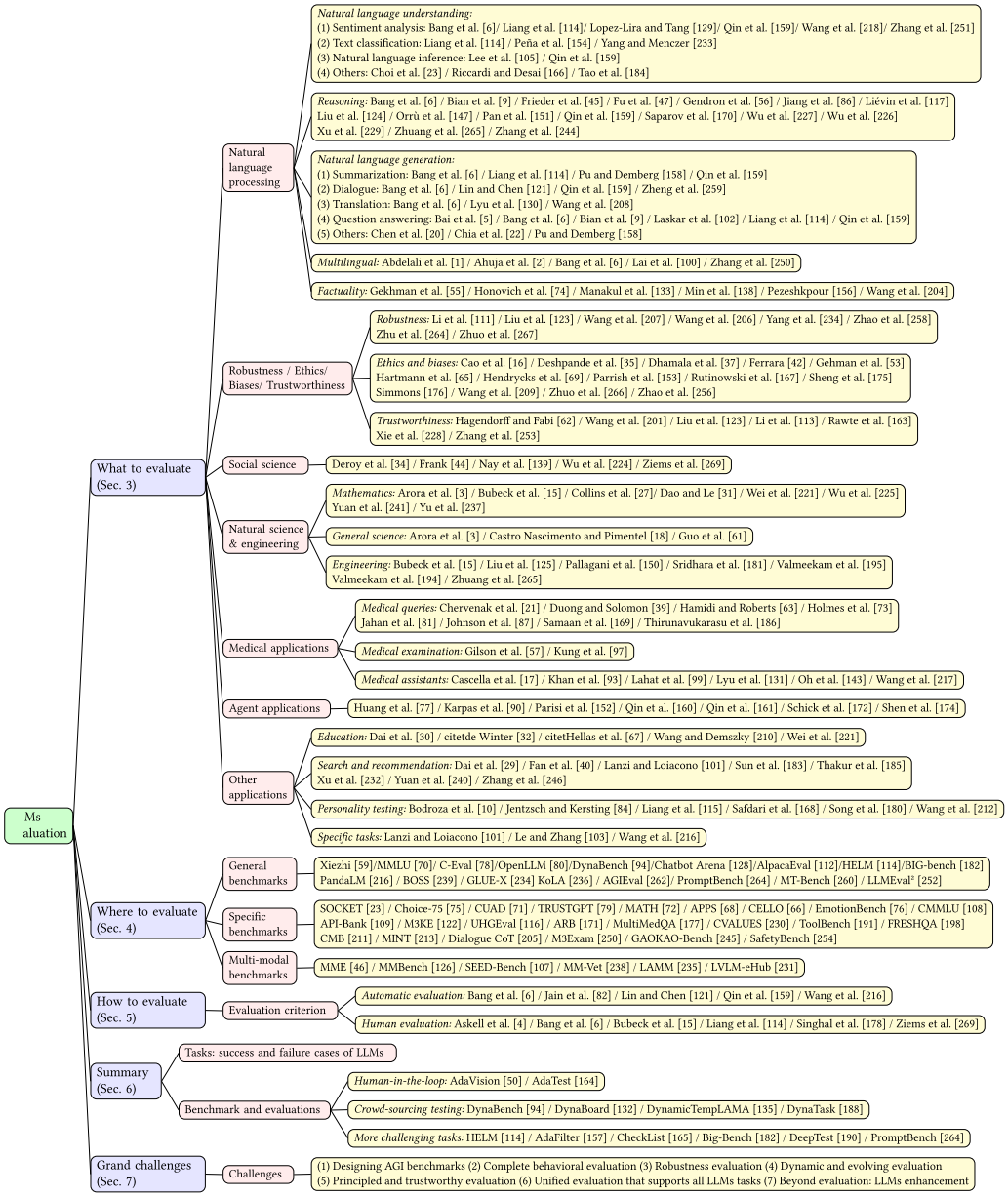

Fig. 1. Structure of this paper.

also face challenges, such as the issue of rare or unseen words, the problem of overfitting, and the difficulty in capturing complex linguistic phenomena. Researchers are continuously working on improving LM architectures and training methods to address these challenges.

Large Language Models (LLMs) [19, 91, 255] are advanced language models with massive parameter sizes and exceptional learning capabilities. The core module behind many LLMs such as GPT-3 [43], InstructGPT [149], and GPT-4 [146] is the self-attention module in Transformer [197]

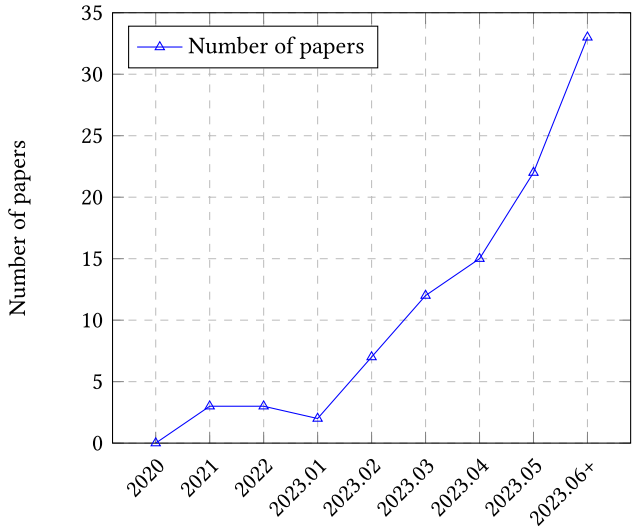

Fig. 2. Trend of LLMs evaluation papers over time (2020 - Jun. 2023, including Jul. 2023.)

that serves as the fundamental building block for language modeling tasks. Transformers have revolutionized the field of NLP with their ability to handle sequential data efficiently, allowing for parallelization and capturing long-range dependencies in text. One key feature of LLMs is in-context learning [14], where the model is trained to generate text based on a given context or prompt. This enables LLMs to generate more coherent and contextually relevant responses, making them suitable for interactive and conversational applications. **Reinforcement Learning from Human Feedback (RLHF)** [25, 266] is another crucial aspect of LLMs. This technique involves fine-tuning the model using human-generated responses as rewards, allowing the model to learn from its mistakes and improve its performance over time.

In an autoregressive language model, such as GPT-3 and PaLM [24], given a context sequence X, the LM tasks aim to predict the next token y. The model is trained by maximizing the probability of the given token sequence conditioned on the context, i.e., $P(y|X) = P(y|x_1, x_2, \ldots, x_{t-1})$, where $x_1, x_2, \ldots, x_{t-1}$ are the tokens in the context sequence, and t is the current position. By using the chain rule, the conditional probability can be decomposed into a product of probabilities at each position:

$$P(y|X) = \prod_{t=1}^{T} P(y_t|x_1, x_2, \ldots, x_{t-1}),$$

where T is sequence length. In this way, the model predicts each token at each position in an autoregressive manner, generating a complete text sequence.

One common approach to interacting with LLMs is prompt engineering [26, 221, 261], where users design and provide specific prompt texts to guide LLMs in generating desired responses or completing specific tasks. This is widely adopted in existing evaluation efforts. People can also engage in question-and-answer interactions [83], where they pose questions to the model and receive answers, or engage in dialogue interactions, having natural language conversations with LLMs. In conclusion, LLMs, with their Transformer architecture, in-context learning, and RLHF capabilities, have revolutionized NLP and hold promise in various applications. Table 1 provides a brief comparison of traditional ML, deep learning, and LLMs.

Table 1. Comparison of Traditional ML, Deep Learning, and LLMs

Comparison	Traditional ML	Deep Learning	LLMs
Training Data Size	Large	Large	Very large
Feature Engineering	Manual	Automatic	Automatic
Model Complexity	Limited	Complex	Very Complex
Interpretability	Good	Poor	Poorer
Performance	Moderate	High	Highest
Hardware Requirements	Low	High	Very High

Fig. 3. The evaluation process of AI models.

2.2 AI Model Evaluation

AI model evaluation is an essential step in assessing the performance of a model. There are some standard model evaluation protocols, including k-fold cross-validation, holdout validation, **leave one out cross-validation (LOOCV)**, bootstrap, and reduced set [8, 95]. For instance, k-fold cross-validation divides the dataset into k parts, with one part used as a test set and the rest as training sets, which can reduce training data loss and obtain relatively more accurate model performance evaluation [48]; Holdout validation divides the dataset into training and test sets, with a smaller calculation amount but potentially more significant bias; LOOCV is a unique k-fold cross-validation method where only one data point is used as the test set [222]; Reduced set trains the model with one dataset and tests it with the remaining data, which is computationally simple, but the applicability is limited. The appropriate evaluation method should be chosen according to the specific problem and data characteristics for more reliable performance indicators.

Figure 3 illustrates the evaluation process of AI models, including LLMs. Some evaluation protocols may not be feasible to evaluate deep learning models due to the extensive training size. Thus, evaluation on a static validation set has long been the standard choice for deep learning models. For instance, computer vision models leverage static test sets such as ImageNet [33] and MS COCO [120] for evaluation. LLMs also use GLUE [200] or SuperGLUE [199] as the common test sets.

As LLMs are becoming more popular with even poorer interpretability, existing evaluation protocols may not be enough to evaluate the true capabilities of LLMs thoroughly. We will introduce recent evaluations of LLMs in Section 5.

3 WHAT TO EVALUATE

What tasks should we evaluate on LLMs to show their performance? On what tasks can we claim the strengths and weaknesses of LLMs? In this section, we divide existing tasks into the following categories: natural language processing, robustness, ethics, biases and trustworthiness, social sciences, natural science and engineering, medical applications, agent applications (using LLMs as agents), and other applications.[1]

[1]Note that LLMs are evaluated in various tasks and the categorization in this paper is only one possible way for classification of these works. There are certainly other taxonomies.

3.1 Natural Language Processing Tasks

The initial objective behind the development of language models, particularly large language models, was to enhance performance on natural language processing tasks, encompassing both understanding and generation. Consequently, the majority of evaluation research has been primarily focused on natural language tasks. Table 2 summarizes the evaluation aspects of existing research, and we mainly highlight their conclusions in the following.[2]

3.1.1 Natural Language Understanding. Natural language understanding represents a wide spectrum of tasks that aims to obtain a better understanding of the input sequence. We summarize recent efforts in LLMs evaluation from several aspects.

Sentiment analysis is a task that analyzes and interprets the text to determine the emotional inclination. It is typically a binary (positive and negative) or triple (positive, neutral, and negative) class classification problem. Evaluating sentiment analysis tasks is a popular direction. Liang et al. [114] and Zeng et al. [242] showed that the performance of the models on this task is usually high. ChatGPT's sentiment analysis prediction performance is superior to traditional sentiment analysis methods [129] and comes close to that of GPT-3.5 [159]. In fine-grained sentiment and emotion cause analysis, ChatGPT also exhibits exceptional performance [218]. In low-resource learning environments, LLMs exhibit significant advantages over small language models [249], but the ability of ChatGPT to understand low-resource languages is limited [6]. In conclusion, LLMs have demonstrated commendable performance in sentiment analysis tasks. Future work should focus on enhancing their capability to understand emotions in under-resourced languages.

Text classification and sentiment analysis are related fields; text classification not only focuses on sentiment, but also includes the processing of all texts and tasks. The work of Liang et al. [114] showed that GLM-130B was the best-performed model, with an overall accuracy of 85.8% for miscellaneous text classification. Yang and Menczer [232] found that ChatGPT can produce credibility ratings for a wide range of news outlets, and these ratings have a moderate correlation with those from human experts. Furthermore, ChatGPT achieves acceptable accuracy in a binary classification scenario (AUC=0.89). Peña et al. [154] discussed the problem of topic classification for public affairs documents and showed that using an LLM backbone in combination with SVM classifiers is a useful strategy to conduct the multi-label topic classification task in the domain of public affairs with accuracies over 85%. Overall, LLMs perform well on text classification and can even handle text classification tasks in unconventional problem settings as well.

Natural language inference (NLI) is the task of determining whether the given "hypothesis" logically follows from the "premise". Qin et al. [159] showed that ChatGPT outperforms GPT-3.5 for NLI tasks. They also found that ChatGPT excels in handling factual input that could be attributed to its RLHF training process in favoring human feedback. However, Lee et al. [105] observed LLMs perform poorly in the scope of NLI and further fail in representing human disagreement, which indicates that LLMs still have large room for improvement in this field.

Semantic understanding refers to the meaning or understanding of language and its associated concepts. It involves the interpretation and comprehension of words, phrases, sentences, and the relationships between them. Semantic processing goes beyond the surface level and focuses on understanding the underlying meaning and intent. Tao et al. [184] comprehensively evaluated the event semantic processing abilities of LLMs covering understanding, reasoning, and prediction about the event semantics. Results indicated that LLMs possess an understanding of individual events, but their capacity to perceive the semantic similarity among events is constrained. In reasoning tasks, LLMs exhibit robust reasoning abilities in causal and intentional relations, yet their

[2]Several NLP areas have intersections and thus our categorization of these areas is only one possible way to categorize.

Table 2. Summary of Evaluation on **Natural Language Processing** Tasks: **NLU (Natural Language Understanding**, Including **SA (Sentiment Analysis)**, **TC (Text Classification)**, **NLI (Natural Language Inference)** and other NLU Tasks), Reasoning, **NLG (Natural Language Generation**, Including Summ

Reference	NLU				RNG.	NLG					Mul.
	SA	TC	NLI	Others		Summ.	Dlg.	Tran.	QA	Others	
Abdelali et al. [1]											✓
Ahuja et al. [2]											✓
Bian et al. [9]					✓				✓		
Bang et al. [6]	✓				✓	✓	✓	✓	✓		✓
Bai et al. [5]									✓		
Chen et al. [20]										✓	
Choi et al. [23]				✓							
Chia et al. [22]										✓	
Frieder et al. [45]					✓						
Fu et al. [47]					✓						
Gekhman et al. [55]						✓					
Gendron et al. [56]					✓						
Honovich et al. [74]			✓			✓	✓			✓	
Jiang et al. [86]					✓						
Lai et al. [100]											✓
Laskar et al. [102]	✓		✓		✓	✓		✓	✓	✓	✓
Lopez-Lira and Tang [129]	✓										
Liang et al. [114]	✓	✓				✓			✓		
Lee et al. [105]			✓								
Lin and Chen [121]							✓				
Liévin et al. [117]					✓						
Liu et al. [124]					✓						
Lyu et al. [130]									✓		
Manakul et al. [133]									✓	✓	
Min et al. [138]										✓	
Orrù et al. [147]					✓						
Pan et al. [151]					✓						
Peña et al. [154]		✓									
Pu and Demberg [158]						✓				✓	
Pezeshkpour [156]										✓	
Qin et al. [159]	✓		✓		✓	✓	✓		✓		
Riccardi and Desai [166]				✓							
Saparov et al. [170]					✓						
Tao et al. [184]				✓							
Wang et al. [208]								✓			
Wang et al. [218]	✓										
Wang et al. [204]			✓						✓		
Wu et al. [226]					✓						
Wu et al. [225]					✓						
Xu et al. [228]					✓						
Yang and Menczer [232]		✓									
Zheng et al. [257]							✓				
Zhang et al. [249]	✓										
Zhang et al. [248]											✓
Zhuang et al. [263]					✓						
Zhang et al. [243]					✓						

(Summarization), Dlg. (Dialogue), Tran (Translation), QA (Question Answering) and other NLG tasks), and Multilingual tasks (ordered by the name of the first author).

performance in other relation types is comparatively weaker. In prediction tasks, LLMs exhibit enhanced predictive capabilities for future events with increased contextual information. Riccardi and Desai [166] explored the semantic proficiency of LLMs and showed that these models perform poorly in evaluating basic phrases. Furthermore, GPT-3.5 and Bard cannot distinguish between

meaningful and nonsense phrases, consistently classifying highly nonsense phrases as meaningful. GPT-4 shows significant improvements, but its performance is still significantly lower than that of humans. In summary, the performance of LLMs in semantic understanding tasks is poor. In the future, we can start from this aspect and focus on improving its performance on this application.

In **social knowledge understanding**, Choi et al. [23] evaluated how well models perform at learning and recognizing concepts of social knowledge and the results revealed that despite being much smaller in the number of parameters, finetuning supervised models such as BERT lead to much better performance than zero-shot models using state-of-the-art LLMs, such as GPT [162], GPT-J-6B [202] and so on. This statement demonstrates that supervised models significantly outperform zero-shot models in terms of performance, highlighting that an increase in parameters does not necessarily guarantee a higher level of social knowledge in this particular scenario.

3.1.2 Reasoning. The task of reasoning poses significant challenges for an intelligent AI model. To effectively tackle reasoning tasks, the models need to not only comprehend the provided information but also utilize reasoning and inference to deduce answers when explicit responses are absent. Table 2 reveals that there is a growing interest in evaluating the reasoning ability of LLMs, as evidenced by the increasing number of articles focusing on exploring this aspect. Currently, the evaluation of reasoning tasks can be broadly categorized into mathematical reasoning, commonsense reasoning, logical reasoning, and domain-specific reasoning.

ChatGPT exhibits a strong capability for arithmetic reasoning by outperforming GPT-3.5 in the majority of tasks [159]. However, its proficiency in mathematical reasoning still requires improvement [6, 45, 263]. On symbolic reasoning tasks, ChatGPT is mostly worse than GPT-3.5, which may be because ChatGPT is prone to uncertain responses, leading to poor performance [6]. Through the poor performance of LLMs on task variants of counterfactual conditions, Wu et al. [226] showed that the current LLMs have certain limitations in abstract reasoning ability. On abstract reasoning, Gendron et al. [56] found that existing LLMs have very limited ability. In logical reasoning, Liu et al. [124] indicated that ChatGPT and GPT-4 outperform traditional fine-tuning methods on most benchmarks, demonstrating their superiority in logical reasoning. However, both models face challenges when handling new and out-of-distribution data. ChatGPT does not perform as well as other LLMs, including GPT-3.5 and BARD [159, 228]. This is because ChatGPT is designed explicitly for chatting, so it does an excellent job of maintaining rationality. FLAN-T5, LLaMA, GPT-3.5, and PaLM perform well in general deductive reasoning tasks [170]. GPT-3.5 is not good at keeping oriented for reasoning in the inductive setting [228]. For multi-step reasoning, Fu et al. [47] showed PaLM and Claude2 are the only two model families that achieve similar performance (but still worse than the GPT model family). Moreover, LLaMA-65B is the most robust open-source LLMs to date, which performs closely to code-davinci-002. Some papers separately evaluate the performance of ChatGPT on some reasoning tasks: ChatGPT generally performs poorly on commonsense reasoning tasks, but relatively better than non-text semantic reasoning [6]. Meanwhile, ChatGPT also lacks spatial reasoning ability, but exhibits better temporal reasoning. Finally, while the performance of ChatGPT is acceptable on causal and analogical reasoning, it performs poorly on multi-hop reasoning ability, which is similar to the weakness of other LLMs on complex reasoning [148]. In professional domain reasoning tasks, zero-shot InstructGPT and Codex are capable of complex medical reasoning tasks, but still need to be further improved [117]. In terms of language insight issues, Orrù et al. [147] demonstrated the potential of ChatGPT for solving verbal insight problems, as ChatGPT's performance was comparable to that of human participants. It should be noted that most of the above conclusions are obtained for specific data sets. In contrast, more complex tasks have become the mainstream benchmarks for assessing the capabilities of LLMs. These include tasks such as mathematical reasoning [225, 236, 243] and structured

data inference [86, 151]. Overall, LLMs show great potential in reasoning and show a continuous improvement trend, but still face many challenges and limitations, requiring more in-depth research and optimization.

3.1.3 Natural Language Generation. NLG evaluates the capabilities of LLMs in generating specific texts, which consists of several tasks, including summarization, dialogue generation, machine translation, question answering, and other open-ended generation tasks.

Summarization is a generation task that aims to learn a concise abstract for the given sentence. In this evaluation, Liang et al. [114] found that TNLG v2 (530B) [179] achieved the highest score in both scenarios, followed by OPT (175B) [245] in second place. The fine-tuned Bart [106] is still better than zero-shot ChatGPT. Specifically, ChatGPT demonstrates comparable zero-shot performance to the text-davinci-002 [6], but performs worse than GPT-3.5 [159]. These findings indicate that LLMs, particularly ChatGPT, have a general performance in summarization tasks.

Evaluating the performance of LLMs on **dialogue** tasks is crucial to the development of dialogue systems and improving human-computer interaction. Through such evaluation, the natural language processing ability, context understanding ability and generation ability of the model can be improved, so as to realize a more intelligent and more natural dialogue system. Both Claude and ChatGPT generally achieve better performance across all dimensions when compared to GPT-3.5 [121, 159]. When comparing the Claude and ChatGPT models, both models demonstrate competitive performance across different evaluation dimensions, with Claude slightly outperforming ChatGPT in specific configurations. Research by Bang et al. [6] underscores that fully fine-tuned models tailored for specific tasks surpass ChatGPT in both task-oriented and knowledge-based dialogue contexts. Additionally, Zheng et al. [257] have curated a comprehensive LLMs conversation dataset, LMSYS-Chat-1M, encompassing up to one million samples. This dataset serves as a valuable resource for evaluating and advancing dialogue systems.

While LLMs are not explicitly trained for **translation** tasks, they can still demonstrate strong performance. Wang et al. [208] demonstrated that ChatGPT and GPT-4 exhibit superior performance in comparison to commercial **machine translation (MT)** systems, as evaluated by humans. Additionally, they outperform most document-level NMT methods in terms of sacreBLEU scores. During contrastive testing, ChatGPT shows lower accuracy in comparison to traditional translation models. However, GPT-4 demonstrates a robust capability in explaining discourse knowledge, even though it may occasionally select incorrect translation candidates. The findings from Bang et al. [6] indicated that ChatGPT performs X $\rightarrow$ Eng translation well, but it still lacks the ability to perform Eng $\rightarrow$ X translation. Lyu et al. [130] investigated several research directions in MT utilizing LLMs. This study significantly contributes to the advancement of MT research and highlights the potential of LLMs in enhancing translation capabilities. In summary, while LLMs perform satisfactorily in several translation tasks, there is still room for improvement, e.g., enhancing the translation capability from English to non-English languages.

Question answering is a crucial technology in the field of human-computer interaction, and it has found wide application in scenarios like search engines, intelligent customer service, and QA systems. The measurement of accuracy and efficiency in QA models will have significant implications for these applications. According to Liang et al. [114], among all the evaluated models, InstructGPT davinci v2 (175B) exhibited the highest performance in terms of accuracy, robustness, and fairness across the 9 QA scenarios. Both GPT-3.5 and ChatGPT demonstrate significant advancements compared to GPT-3 in their ability to answer general knowledge questions. In most domains, ChatGPT surpasses GPT-3.5 by more than 2% in terms of performance [9, 159]. However, ChatGPT performs slightly weaker than GPT-3.5 on the CommonsenseQA and Social IQA benchmarks. This can be attributed to ChatGPT's cautious nature, as it tends to decline to provide an answer when there is insufficient information available. Fine-tuned models, such as Vícuna and

ChatGPT, exhibit exceptional performance with near-perfect scores, surpassing models that lack supervised fine-tuning by a significant margin [5, 6]. Laskar et al. [102] evaluated the effectiveness of ChatGPT on a range of academic datasets, including various tasks such as answering questions, summarizing text, generating code, reasoning with commonsense, solving math problems, translating languages, detecting bias, and addressing ethical issues. Overall, LLMs showcase flawless performance on QA tasks and hold the potential for further enhancing their proficiency in social, event, and temporal commonsense knowledge in the future.

There are also other generation tasks to explore. In the field of **sentence style transfer**, Pu and Demberg [158] demonstrated that ChatGPT surpasses the previous SOTA supervised model through training on the same subset for few-shot learning, as evident from the higher BLEU score. However, when it comes to controlling the formality of sentence style, ChatGPT's performance still differs significantly from human behavior. In **writing tasks**, Chia et al. [22] discovered that LLMs exhibit consistent performance across various categories such as informative, professional, argumentative, and creative writing. This finding implies that LLMs possess a general proficiency in writing capabilities. In **text generation** quality, Chen et al. [20] revealed that ChatGPT excels in assessing text quality from multiple angles, even in the absence of reference texts, surpassing the performance of most existing automated metrics. Employing ChatGPT to generate numerical scores for text quality emerged as the most reliable and effective approach among the various testing methods studied.

3.1.4 Multilingual Tasks. While English is the predominant language, many LLMs are trained on mixed-language training data. The combination of multilingual data indeed helps LLMs gain the ability to process inputs and generate responses in different languages, making them widely adopted and accepted across the globe. However, due to the relatively recent emergence of this technology, LLMs are primarily evaluated on English data, leading to a potential oversight of evaluating their multilingual performance. To address this, several articles have provided comprehensive, open, and independent evaluations of LLMs' performance on various NLP tasks in different non-English languages. These evaluations offer valuable insights for future research and applications.

Abdelali et al. [1] evaluated the performance of ChatGPT in standard Arabic NLP tasks and observed that ChatGPT exhibits lower performance compared to SOTA models in the zero-shot setting for most tasks. Ahuja et al. [2], Bang et al. [6], Lai et al. [100], Zhang et al. [248] utilized a greater number of languages across multiple datasets, encompassing a wider range of tasks, and conducted a more comprehensive evaluation of LLMs, including BLOOM, Vicuna, Claude, ChatGPT, and GPT-4. The results indicated that these LLMs perform poorly when it came to non-Latin languages and languages with limited resources. Despite translating the input to English and using it as the query, generative LLMs still displays subpar performance across tasks and languages compared to SOTA models [2]. Furthermore, Bang et al. [6] highlighted that ChatGPT still faces a limitation in translating sentences written in non-Latin script languages with rich linguistic resources. The aforementioned demonstrates that there are numerous challenges and ample opportunities for enhancement in multilingual tasks for LLMs. Future research should prioritize achieving multilingual balance and addressing the challenges faced by non-Latin languages and low-resource languages, with the aim of better supporting users worldwide. At the same time, attention should be paid to the impartiality and neutrality of the language in order to mitigate any potential biases, including English bias or other biases, that could impact multilingual applications.

3.1.5 Factuality. Factuality in the context of LLMs refers to the extent to which the information or answers provided by the model align with real-world truths and verifiable facts. Factuality in LLMs significantly impacts a variety of tasks and downstream applications, such as QA systems,

information extraction, text summarization, dialogue systems, and automated fact-checking, where incorrect or inconsistent information could lead to substantial misunderstandings and misinterpretations. Evaluating factuality is of great importance in order to trust and efficiently use these models. This includes the ability of these models to maintain consistency with known facts, avoid generating misleading or false information (known as "factual hallucination"), and effectively learn and recall factual knowledge. A range of methodologies have been proposed to measure and improve the factuality of LLMs.

Wang et al. [204] assessed the internal knowledge capabilities of several large models, namely InstructGPT, ChatGPT-3.5, GPT-4, and BingChat [137], by examining their ability to answer open questions based on the Natural Questions [98] and TriviaQA [88] datasets. The evaluation process involved human assessment. The results of the study indicated that while GPT-4 and BingChat can provide correct answers for more than 80% of the questions, there is still a remaining gap of over 15% to achieve complete accuracy. In the work of Honovich et al. [74], they conducted a review of current factual consistency evaluation methods and highlighted the absence of a unified comparison framework and the limited reference value of related scores compared to binary labels. To address this, they transformed existing fact consistency tasks into binary labels, specifically considering only whether there is a factual conflict with the input text, without factoring in external knowledge. The research discovered that fact evaluation methods founded on natural language inference and question generation answering exhibit superior performance and can complement each other. Pezeshkpour [156] proposed a novel metric, based on information theory, to assess the inclusion of specific knowledge in LLMs. The metric utilized the concept of uncertainty in knowledge to measure factualness, calculated by LLMs filling in prompts and examining the probability distribution of the answer. The paper discussed two methods for injecting knowledge into LLMs: explicit inclusion of knowledge in the prompts and implicit fine-tuning of the LLMs using knowledge-related data. The study demonstrated that this approach surpasses traditional ranking methods by achieving an accuracy improvement of over 30%. Gekhman et al. [55] improved the method for evaluating fact consistency in summarization tasks. It proposed a novel approach that involved training student NLI models using summaries generated by multiple models and annotated by LLMs to ensure fact consistency. The trained student model was then used for summarization fact consistency evaluation. Manakul et al. [133] operated on two hypotheses regarding how LLMs generate factual or hallucinated responses. It proposed the use of three formulas (BERTScore [247], MQAG [134] and n-gram) to evaluate factuality and employed alternative LLMs to gather token probabilities for black-box language models. The study discovered that simply computing sentence likelihood or entropy helped validate the factuality of the responses. Min et al. [138] broke down text generated by LLMs into individual "atomic" facts, which were then evaluated for their correctness. The FActScore is used to measure the performance of estimators through the calculation of F1 scores. The paper tested various estimators and revealed that current estimators still have some way to go in effectively addressing the task. Lin et al. [119] introduced the TruthfulQA dataset, designed to cause models to make mistakes. Multiple language models were tested by providing factual answers. The findings from these experiments suggest that simply scaling up model sizes may not necessarily improve their truthfulness, and recommendations are provided for the training approach. This dataset has become widely used for evaluating the factuality of LLMs [89, 146, 192, 219].

3.2 Robustness, Ethics, Bias, and Trustworthiness

The evaluation encompasses crucial aspects of robustness, ethics, biases, and trustworthiness. These factors have gained increasing importance in assessing the performance of LLMs comprehensively. Table 3 shows a summary of the research.

Table 3. Summary of LLMs Evaluation on **Robustness, Ethics, Biases, and Trustworthiness** (Ordered by the Name of the First Author)

Reference	Robustness	Ethics and biases	Trustworthiness
Cao et al. [16]		✓	
Dhamala et al. [37]		✓	
Deshpande et al. [35]		✓	
Ferrara [42]		✓	
Gehman et al. [53]		✓	
Hartmann et al. [65]		✓	
Hendrycks et al. [69]		✓	
Hagendorff and Fabi [62]			✓
Li et al. [111]	✓		
Liu et al. [123]	✓		
Liu et al. [123]			✓
Li et al. [113]			✓
Parrish et al. [153]		✓	
Rutinowski et al. [167]		✓	
Rawte et al. [163]			✓
Sheng et al. [175]		✓	
Simmons [176]		✓	
Wang et al. [207]	✓		
Wang et al. [206]	✓		
Wang et al. [201]	✓	✓	✓
Wang et al. [209]		✓	
Xie et al. [227]			✓
Yang et al. [233]	✓		
Zhao et al. [256]	✓		
Zhuo et al. [265]	✓		
Zhu et al. [262]	✓		
Zhuo et al. [264]		✓	
Zhang et al. [251]			✓

3.2.1 Robustness. Robustness studies the stability of a system when facing unexpected inputs. Specifically, **out-of-distribution (OOD)** [207] and adversarial robustness are two popular research topics for robustness. Wang et al. [206] is an early work that evaluated ChatGPT and other LLMs from both the adversarial and OOD perspectives using existing benchmarks such as AdvGLUE [203], ANLI [140], and DDXPlus [41] datasets. Zhuo et al. [265] evaluated the robustness of semantic parsing. Yang et al. [233] evaluated OOD robustness by extending the GLUE [200] dataset. The results of this study emphasize the potential risks to the overall system security when manipulating visual input. For vision-language models, Zhao et al. [256] evaluated LLMs on visual input and transferred them to other visual-linguistic models, revealing the vulnerability of visual input. Li et al. [111] provided an overview of OOD evaluation for language models: adversarial robustness, domain generalization, and dataset biases. Bridging these lines of research, the authors conducted a comparative analysis, unifying the three approaches. They succinctly outlined the data-generation processes and evaluation protocols for each line of study, all while emphasizing the prevailing challenges and future research prospects. Additionally, Liu et al. [123] introduced a large-scale robust visual instruction dataset to enhance the performance of large-scale multi-modal models in handling relevant images and human instructions.

For adversarial robustness, Zhu et al. [262] evaluated the robustness of LLMs to prompts by proposing a unified benchmark called PromptBench. They comprehensively evaluated adversarial text attacks at multiple levels (character, word, sentence, and semantics). The results showed that contemporary LLMs are vulnerable to adversarial prompts, highlighting the importance of

the models' robustness when facing adversarial inputs. As for new adversarial datasets, Wang et al. [201] introduced AdvGLUE++ benchmark data for assessing adversarial robustness and implemented a new evaluation protocol to scrutinize machine ethics via jailbreaking system prompts.

3.2.2 Ethics and Bias. LLMs have been found to internalize, spread, and potentially magnify harmful information existing in the crawled training corpora, usually, toxic languages, like offensiveness, hate speech, and insults [53], as well as social biases like stereotypes towards people with a particular demographic identity (e.g., gender, race, religion, occupation, and ideology) [175]. More recently, Zhuo et al. [264] used conventional testing sets and metrics [37, 53, 153] to perform a systematic evaluation of ChatGPT's toxicity and social bias, finding that it still exhibits noxious content to some extend. Taking a further step, Deshpande et al. [35] introduced role-playing into the model and observed an increase in generated toxicity up to 6x. Furthermore, such role-playing also caused biased toxicity towards specific entities. Different from simply measuring social biases, Ferrara [42] investigated the sources, underlying mechanisms, and corresponding ethical consequences of these biases potentially produced by ChatGPT. Beyond social biases, LLMs have also been assessed by political tendency and personality traits [65, 167] based questionnaires like the Political Compass Test and MBTI test, demonstrating a propensity for progressive views and an ENFJ personality type. In addition, LLMs like GPT-3 were found to have moral biases [176] in terms of the Moral Foundation theory [58]; The study conducted by [69] reveals that existing LMs have potential in ethical judgment, but still need improvement. [254] proposes a Chinese conversational bias evaluation dataset CHBias, discovers bias risks in pre-trained models, and explores debiasing methods. Moreover, in the assessment of GPT-4 alignment, [209] discovered a systematic bias. ChatGPT is also observed to exhibit somewhat bias on cultural values [16]. Wang et al. [201] also incorporated an evaluation dataset specifically aimed at gauging stereotype bias, using both targeted and untargeted system prompts. All these ethical issues might elicit serious risks, impeding the deployment of LLMs and having a profound negative impact on society.

3.2.3 Trustworthiness. Some work focuses on other trustworthiness problems in addition to robustness and ethics.[3] In their 2023 study, DecodingTrust, Wang et al. [201] offered a multifaceted exploration of trustworthiness vulnerabilities in the GPT models, especially GPT-3.5 and GPT-4. Their evaluation expanded beyond the typical trustworthiness concerns to include eight critical aspects: toxicity, stereotype bias, adversarial and out-of-distribution robustness, robustness to adversarial demonstrations, privacy, machine ethics, and fairness. DecodingTrust's investigation employs an array of newly constructed scenarios, tasks, and metrics. They revealed that while GPT-4 often showcases improved trustworthiness over GPT-3.5 in standard evaluations, it is simultaneously more susceptible to attacks.

In another study by Hagendorff and Fabi [62], LLMs with enhanced cognitive abilities were evaluated. They found that these models can avoid common human intuitions and cognitive errors, demonstrating super-rational performance. By utilizing cognitive reflection tests and semantic illusion experiments, the researchers gained insights into the psychological aspects of LLMs. This method offers new perspectives for evaluating model biases and ethical issues that may not have been previously identified. Furthermore, a study by [227] brings attention to a significant concern: the consistency of judgment in LLMs diminishes notably when faced with disruptions such as questioning, negation, or misleading cues, even if their initial judgments were accurate. The research delves into various prompting methods designed to mitigate this issue and successfully demonstrates their efficacy.

[3]The term 'trustworthiness' in this section refers to other work that contains more than robustness and ethics.

LLMs are capable of generating coherent and seemingly factual text. However, the information generated can include factual inaccuracies or statements ungrounded in reality, a phenomenon known as **hallucination** [163, 251]. Evaluating these issues helps improve the training methods of LLMs to reduce the occurrence of hallucinations. For the evaluation of illusions in large-scale visual models, Liu et al. [123] introduced a comprehensive and robust large-scale visual instruction dataset: LRV-Instruction. Through the GAVIE method, they fine-tuned the evaluation visual instructions, and experimental results demonstrated that LRV-Instruction effectively alleviates illusions in LLMs. In addition, Li et al. [113] conducted an assessment of illusions in large-scale visual language models, revealing through experiments that the distribution of objects in visual instructions significantly impacts object illusions in LVLMs. To enhance the assessment of object illusions in LVLMs, they introduced a polling-based query method, known as POPE. This method provides an improved evaluation of object illusions in LVLMs.

3.3 Social Science

Social science involves the study of human society and individual behavior, including economics, sociology, political science, law, and other disciplines. Evaluating the performance of LLMs in social science is important for academic research, policy formulation, and social problem-solving. Such evaluations can help improve the applicability and quality of models in the social sciences, increasing understanding of human societies and promoting social progress.

Wu et al. [223] evaluated the potential use of LLMs in addressing scaling and measurement issues in social science and found that LLMs can generate meaningful responses regarding political ideology and significantly improve text-as-data methods in social science.

In **computational social science (CSS)** tasks, Ziems et al. [267] presented a comprehensive evaluation of LLMs on several CSS tasks. During classification tasks, LLMs exhibit the lowest absolute performance on event argument extraction, character tropes, implicit hate, and empathy classification, achieving accuracy below 40%. These tasks either involve complex structures (event arguments) or subjective expert taxonomies with semantics that differ from those learned during LLM pretraining. Conversely, LLMs achieve the best performance on misinformation, stance, and emotion classification. When it comes to generation tasks, LLMs often produce explanations that surpass the quality of gold references provided by crowd workers. In summary, while LLMs can greatly enhance the traditional CSS research pipeline, they cannot completely replace it.

Some articles also evaluate LLMs on legal tasks. The zero-shot performance of LLMs is mediocre in legal case judgment summarization. LLMs have several problems, including incomplete sentences and words, meaningless sentences merge, and more serious errors such as inconsistent and hallucinated information [34]. The results showed that further improvement is necessary for LLMs to be useful for case judgment summarization by legal experts. Nay et al. [139] indicated that LLMs, particularly when combined with prompting enhancements and the correct legal texts, could perform better but not yet at expert tax lawyer levels.

Lastly, within the realm of psychology, Frank [44] adopted an interdisciplinary approach and drew insights from developmental psychology and comparative psychology to explore alternative methods for evaluating the capabilities of LLMs. By integrating different perspectives, researchers can deepen their understanding of the essence of cognition and effectively leverage the potential of advanced technologies such as large language models, while mitigating potential risks.

In conclusion, the utilization of LLMs has significantly benefited individuals in addressing social science-related tasks, leading to improved work efficiency. The outputs produced by LLMs serve as valuable resources for enhancing productivity. However, it is crucial to acknowledge that existing LLMs cannot completely replace human professionals in this domain.

Table 4. Summary of Evaluations on **Natural Science and Engineering Tasks** Based on Three Aspects: Mathematics, General Science and Engineering (Ordered by the Name of the First Author)

Reference	Mathematics	General science	Engineering
Arora et al. [3]	✓	✓	
Bubeck et al. [15]	✓		✓
Castro Nascimento and Pimentel [18]		✓	
Collins et al. [27]	✓		
Dao and Le [31]	✓		
Guo et al. [61]		✓	
Liu et al. [125]			✓
Pallagani et al. [150]			✓
Sridhara et al. [181]			✓
Valmeekam et al. [194]			✓
Valmeekam et al. [195]			✓
Wei et al. [220]	✓		
Wu et al. [224]	✓		
Yuan et al. [240]	✓		
Yu et al. [236]	✓		
Zhuang et al. [263]			✓

3.4 Natural Science and Engineering

Evaluating the performance of LLMs in natural science and engineering can help guide applications and development in scientific research, technology development, and engineering studies. Table 4 shows a summary of the natural science and engineering tasks.

3.4.1 Mathematics. For fundamental mathematical problems, most large language models (LLMs) demonstrate proficiency in addition and subtraction, and possess some capability in multiplication. However, they face challenges when it comes to division, exponentiation, trigonometry functions, and logarithm functions. On the other hand, LLMs exhibit competence in handling decimal numbers, negative numbers, and irrational numbers [240]. In terms of performance, ChatGPT and GPT-4 outperform other models significantly, showcasing their superiority in solving mathematical tasks [220]. These two models have a distinct advantage in dealing with large numbers (greater than 1e12) and complex, lengthy mathematical queries. GPT-4 outperforms ChatGPT by achieving a significant increase in accuracy of 10 percentage points and a reduction in relative error by 50%, due to its superior division and trigonometry abilities, proper understanding of irrational numbers, and consistent step-by-step calculation of long expressions.

When confronted with complex and challenging mathematical problems, LLMs exhibit subpar performance. Specifically, GPT-3 demonstrates nearly random performance, while GPT-3.5 shows improvement, and GPT-4 performs the best [3]. Despite the advancements made in the new models, it is important to note that the peak performance remains relatively low compared to that of experts and these models lack the capability to engage in mathematical research [15]. The specific tasks of algebraic manipulation and calculation continue to pose challenges for GPTs [15, 27]. The primary reasons behind GPT-4's low performance in these tasks are errors in algebraic manipulation and difficulties in retrieving pertinent domain-specific concepts. Wu et al. [224] evaluated the use of GPT-4 on difficult high school competition problems and GPT-4 reached 60% accuracy on half of the categories. Intermediate algebra and precalculus can only be solved with a low accuracy rate of around 20%. ChatGPT is not good at answering questions on topics including derivatives and applications, Oxyz spatial calculus, and spatial geometry [31]. Dao and Le [31], Wei et al. [220] showed that ChatGPT's performance worsens as task difficulty increases: it correctly answered 83% of the questions at the recognition level, 62% at the comprehension level, 27% at the application

level, and only 10% at the highest cognitive complexity level. Given those problems at higher knowledge levels tend to be more complex, requiring in-depth understanding and problem-solving skills, such results are to be expected.

These results indicate that the effectiveness of LLMs is highly influenced by the complexity of problems they encounter. This finding holds significant implications for the design and development of optimized artificial intelligence systems capable of successfully handling these challenging tasks.

3.4.2 General Science. Further improvements are needed in the application of LLMs in the field of chemistry. Castro Nascimento and Pimentel [18] presented five straightforward tasks from various subareas of chemistry to assess ChatGPT's comprehension of the subject, with accuracy ranging from 25% to 100%. Guo et al. [61] created a comprehensive benchmark that encompasses eight practical chemistry tasks, which is designed to assess the performance of LLMs (including GPT-4, GPT-3.5, and Davinci-003) for each chemistry task. Based on the experiment results, GPT-4 demonstrates superior performance compared to the other two models. [3] showed that LLMs perform worse on physics problems than chemistry problems, probably because chemistry problems have lower inference complexity than physics problems in this setting. There are limited evaluation studies on LLMs in the field of general science, and the current findings indicate that further improvement is needed in the performance of LLMs within this domain.

3.4.3 Engineering. Within engineering, the tasks can be organized in ascending order of difficulty, including code generation, software engineering, and commonsense planning.

In code generation tasks, the smaller LLMs trained for the tasks are competitive in performance, and CodeGen-16B [141] is comparable in performance to ChatGPT using a larger parameter setting, reaching about a 78% match [125]. Despite facing challenges in mastering and comprehending certain fundamental concepts in programming languages, ChatGPT showcases a commendable level of coding level [263]. Specifically, ChatGPT has developed superior skills in dynamic programming, greedy algorithm, and search, surpassing highly capable college students, but it struggles in data structure, tree, and graph theory. GPT-4 demonstrates an advanced ability to generate code based on given instructions, comprehend existing code, reason about code execution, simulate the impact of instructions, articulate outcomes in natural language, and execute pseudocode effectively [15].

In software engineering tasks, ChatGPT generally performs well and provides detailed responses, often surpassing both human expert output and SOTA output. However, for certain tasks such as code vulnerability detection and information retrieval-based test prioritization, the current version of ChatGPT fails to provide accurate answers, rendering it unsuitable for these specific tasks [181].

In commonsense planning tasks, LLMs may not perform well, even in simple planning tasks where humans excel [194, 195]. Pallagani et al. [150] demonstrated that the fine-tuned CodeT5 [214] performs the best across all considered domains, with the shortest inference time. Moreover, it explored the capability of LLMs for plan generalization and found that their generalization capabilities appear to be limited. It turns out that LLMs can handle simple engineering tasks, but they perform poorly on complex engineering tasks.

3.5 Medical Applications

The application of LLMs in the medical field has recently received significant attention. As a result, this section aims to provide a comprehensive review of the ongoing efforts dedicated to implementing LLMs in medical applications. We have categorized these applications into three aspects as shown in Table 5: medical query, medical examination, and medical assistants. A detailed

Table 5. Summary of Evaluations on **Medical Applications** based on the Three Aspects: Medical Queries, Medical Assistants, and Medical Examination (Ordered by the Name of the First Author)

Reference	Medical queries	Medical examination	Medical assistants
Cascella et al. [17]			✓
Chervenak et al. [21]	✓		
Duong and Solomon [39]	✓		
Gilson et al. [57]		✓	
Hamidi and Roberts [63]	✓		
Holmes et al. [73]	✓		
Jahan et al. [81]	✓		
Johnson et al. [87]	✓		
Khan et al. [93]			✓
Kung et al. [97]		✓	
Lahat et al. [99]			✓
Lyu et al. [131]			✓
Oh et al. [143]			✓
Samaan et al. [169]	✓		
Thirunavukarasu et al. [186]	✓		
Wang et al. [217]			✓

examination of these categories will enhance our understanding of the potential impact and advantages that LLMs can bring to the medical domain.

3.5.1 Medical Queries. The significance of evaluating LLMs on medical queries lies in providing accurate and reliable medical answers to meet the needs of healthcare professionals and patients for high-quality medical information. As shown in Table 5, the majority of LLMs evaluations in the medical field concentrate on medical queries. ChatGPT generated relatively accurate information for various medical queries, including genetics [39], radiation oncology physics [73], biomedicine [81], and many other medical disciplines [63, 87, 169], demonstrating its effectiveness in the field of medical queries to a certain extent. As for the limitations, Thirunavukarasu et al. [186] assessed ChatGPT's performance in primary care and found that its average score in the student comprehensive assessment falls below the passing score, indicating room for improvement. Chervenak et al. [21] highlighted that while ChatGPT can generate responses similar to existing sources in fertility-related clinical prompts, its limitations in reliably citing sources and potential for fabricating information restrict its clinical utility.

3.5.2 Medical Examination. The studies by Gilson et al. [57], Kung et al. [97] have evaluated the performance of LLMs in medical examination assessment through the **United States Medical Licensing Examination (USMLE)**.[4] In the study of [57], ChatGPT's performance in answering USMLE Step 1 and Step 2 exam questions was assessed using novel multiple-choice question sets. The results indicated that ChatGPT achieves varying accuracies across different datasets. However, the presence of out-of-context information was found to be lower compared to the correct answer in the NBME-Free-Step1 and NBME-Free-Step2 datasets. Kung et al. [97] showed that ChatGPT achieves or approaches the passing threshold in these exams with no tailored training. The model demonstrates high consistency and insight, indicating its potential to assist in medical education and clinical decision-making. ChatGPT can be used as a tool to answer medical questions, provide explanations, and support decision-making processes. This offers additional resources and support for medical students and clinicians in their educational and clinical practices. Moreover,

[4] https://www.usmle.org/

Sharma et al. [173] found that answers generated by ChatGPT are more context-aware with better deductive reasoning abilities compared to Google search results.

3.5.3 Medical Assistants. In the field of medical assistance, LLMs demonstrate potential applications, including research on identifying gastrointestinal diseases [99], dementia diagnosis [217], accelerating the evaluation of COVID-19 literature [93], and their overall potential in healthcare [17]. However, there are also limitations and challenges, such as lack of originality, high input requirements, resource constraints, uncertainty in answers, and potential risks related to misdiagnosis and patient privacy issues.

Moreover, several studies have evaluated the performance and feasibility of ChatGPT in the medical education field. In the study by Oh et al. [143], ChatGPT, specifically GPT-3.5 and GPT-4 models, were evaluated in terms of their understanding of surgical clinical information and their potential impact on surgical education and training. The results indicate an overall accuracy of 46.8% for GPT-3.5 and 76.4% for GPT-4, demonstrating a significant performance difference between the two models. Notably, GPT-4 consistently performs well across different subspecialties, suggesting its capability to comprehend complex clinical information and enhance surgical education and training. Another study by Lyu et al. [131] explores the feasibility of utilizing ChatGPT in clinical education, particularly in translating radiology reports into easily understandable language. The findings demonstrate that ChatGPT effectively translates radiology reports into accessible language and provides general recommendations. Furthermore, the quality of ChatGPT has shown improvement compared to GPT-4. These findings suggest that employing LLMs in clinical education is feasible, although further efforts are needed to address limitations and unlock their full potential.

3.6 Agent Applications

Instead of focusing solely on general language tasks, LLMs can be utilized as powerful tools in various domains. Equipping LLMs with external tools can greatly expand the capabilities of the model [160]. ToolLLM [161] provides a comprehensive framework to equip open-source large language models with tool use capabilities. Huang et al. [77] introduced KOSMOS-1, which is capable of understanding general patterns, following instructions, and learning based on context. The studyof MRKL by Karpas et al. [90] emphasized the importance of understanding when and how to utilize external symbolic tools, as this knowledge is dependent on the capabilities of LLMs, particularly when these tools can reliably perform functions. Additionally, two other studies, Toolformer [172] and TALM [152], explored the utilization of tools to enhance language models. Toolformer employs a training approach to determine the optimal usage of specific APIs and integrates the obtained results into subsequent token predictions. On the other hand, TALM combines indistinguishable tools with text-based methods to augment language models and employs an iterative technique known as "self-play", guided by minimal tool demonstrations. Furthermore, Shen et al. [174] proposed the HuggingGPT framework, which leverages LLMs to connect various AI models within the machine learning community (such as Hugging Face), aiming to address AI tasks.

3.7 Other Applications

In addition to above areas, there have been evaluations in various other domains, including education, search and recommendation, personality testing, and specific applications. Table 6 shows a sumary of these applications.

3.7.1 Education. LLMs have shown promise in revolutionizing the field of education. They have the potential to make significant contributions in several areas, such as assisting students in improving their writing skills, facilitating better comprehension of complex concepts, expediting the delivery of information, and providing personalized feedback to enhance student engagement.

Table 6. Summary of Evaluations on **other Applications** based on the Four Aspects: Education, Search and Recommendation, Personality Testing and Specific Applications (Ordered by the Name of the First Author)

Reference	Education	Search and recommendation	Personality testing	Specific applications
Bodroza et al. [10]			✓	
Dai et al. [30]	✓			
de Winter [32]	✓			
Dai et al. [29]		✓		
Fan et al. [40]		✓		
Hellas et al. [67]	✓			
Jentzsch and Kersting [84]			✓	
Lanzi and Loiacono [101]				✓
Le and Zhang [103]				✓
Li et al. [110]		✓		
Liang et al. [115]			✓	
Sun et al. [183]		✓		
Song et al. [180]			✓	
Safdari et al. [168]			✓	
Thakur et al. [185]		✓		
Wang and Demszky [210]	✓			
Wang et al. [212]			✓	
Wang et al. [216]				✓
Xu et al. [231]		✓		
Yuan et al. [239]		✓		
Zhang et al. [244]		✓		

These applications aim to create more efficient and interactive learning experiences, offering students a broader range of educational opportunities. However, to fully harness the potential of LLMs in education, extensive research and ongoing refinement are necessary.

The evaluation of LLMs for **educational assistance** aims to investigate and assess their potential contributions to the field of education. Such evaluations can be conducted from various perspectives. According to Dai et al. [30], ChatGPT demonstrates the ability to generate detailed, fluent, and coherent feedback that surpasses that of human teachers. It can accurately assess student assignments and provide feedback on task completion, thereby assisting in the development of student skills. However, ChatGPT's responses may lack novelty or insightful perspectives regarding teaching improvement [210]. Additionally, the study conducted by Hellas et al. [67] revealed that LLMs can successfully identify at least one actual problem in student code, although instances of misjudgment are also observed. In conclusion, the utilization of LLMs shows promise in addressing program logic issues, although challenges remain in achieving proficiency in output formatting. It is important to note that while these models can provide valuable insights, they may still generate errors similar to those made by students.

In **educational exams**, researchers aim to evaluate the application effectiveness of LLMs, including automatic scoring, question generation, and learning guidance. de Winter [32] showed that ChatGPT achieves an average of 71.8% correctness, which is comparable to the average score of all participating students. Subsequently, the evaluation was conducted using GPT-4, and it achieved a score of 8.33. Furthermore, this evaluation showed the effectiveness of leveraging bootstrapping that combines randomness via the "temperature" parameter in diagnosing incorrect answers. Zhang et al. [246] claimed that GPT-3.5 can solve MIT math and EECS exams with GPT-4 achieving better performance. However, it turned out to be not fair since they accidentally included the correct answers into the prompts.

3.7.2 Search and Recommendation. The assessment of LLMs in search and recommendation can be broadly categorized into two areas. Firstly, in the realm of **information retrieval**, Sun

et al. [183] investigated the effectiveness of generative ranking algorithms, such as ChatGPT and GPT-4, for information retrieval tasks. Experimental results demonstrate that guided ChatGPT and GPT-4 exhibit competitive performance on popular benchmark tests, even outperforming supervised methods. Additionally, the extraction of ChatGPT's ranking functionality into a specialized model shows superior performance when trained on 10K ChatGPT-generated data compared to training on 400K annotated MS MARCO data in the BEIR dataset [185]. Furthermore, Xu et al. [231] conducted a randomized online experiment to investigate the behavioral differences of users when performing information retrieval tasks using search engines and chatbot tools. Participants were divided into two groups: one using tools similar to ChatGPT and the other using tools similar to Google Search. The results show that the ChatGPT group spent less time on all tasks and the difference between these two groups is not significant.

Secondly, moving to the domain of **recommendation systems**, LLMs have emerged as essential components that leverage their natural language processing capabilities to comprehend user preferences, item descriptions, and contextual information [40]. By incorporating LLMs into recommendation pipelines, these systems can offer more accurate and personalized recommendations, thereby improving user experience and overall recommendation quality. However, it is crucial to address the potential risks associated with using LLMs for recommendations. Recent research by Zhang et al. [244] has highlighted the issue of unfair recommendations generated by ChatGPT. This emphasizes the importance of evaluating fairness when employing LLMs in recommendation scenarios. Dai et al. [29] suggest that ChatGPT exhibits strong performance in recommender systems. The use of listwise ranking is found to strike the best balance between cost and performance. Furthermore, ChatGPT shows promise in addressing the cold-start problem and providing interpretable recommendations. Moreover, the research by Yuan et al. [239] and Li et al. [110] demonstrated the promising potential of the **modality-based recommendation model (MoRec)** and **text-based collaborative filtering (TCF)** in recommendation systems.

3.7.3 Personality Testing. Personality testing aims to measure individuals' personality traits and behavioral tendencies, and LLMs as powerful natural language processing models have been widely applied in such tasks.

Research conducted by Bodroza et al. [10] investigated the personality features of using Davinci-003 as a chatbot and found variations in the consistency of its answers, despite exhibiting prosocial characteristics. However, there remains uncertainty regarding whether the chatbot's responses are driven by conscious self-reflection or algorithmic processes. Song et al. [180] examined the manifestation of personality in language models and discovered that many models perform unreliably in self-assessment tests and exhibit inherent biases. Therefore, it is necessary to develop specific machine personality measurement tools to enhance reliability. These studies offer vital insights to better understand LLMs in personality testing. Safdari et al. [168] proposed a comprehensive approach to conduct effective psychometric testing for the personality traits in the text generated by LLMs. In order to evaluate the emotional intelligence of LLMs, Wang et al. [212] developed a new psychometric assessment method. By referencing a framework constructed from over 500 adults, the authors tested various mainstream LLMs. The results showed that most LLMs achieve above-average scores in **emotional quotient (EQ)**, with GPT-4 scoring 117, surpassing 89% of human participants. However, a multivariate pattern analysis indicated that certain LLMs achieve human-level performance without relying on mechanisms resembling those found in humans. This is evident from the distinct differences in the quality of their representational patterns, as compared to humans. Liang et al. [115] employed the word guessing game to evaluate LLMs' language and theory of mind intelligences, a more engaging and interactive assessment method. Jentzsch and Kersting [84] discussed the challenges of incorporating humor into LLMs, particularly ChatGPT.

They found that while ChatGPT demonstrates impressive capabilities in NLP tasks, it falls short in generating humorous responses. This study emphasizes the importance of humor in human communication and the difficulties that LLMs face in capturing the subtleties and context-dependent nature of humor. It discusses the limitations of current approaches and highlights the need for further research on more sophisticated models that can effectively understand and generate humor.

3.7.4 Specific Applications. Moreover, various research endeavors have been conducted to explore the application and evaluation of LLMs across a wide spectrum of tasks, such as **game design** [101], **model performance assessment** [216], and **log parsing** [103]. Collectively, these findings enhance our comprehension of the practical implications associated with the utilization of LLMs across diverse tasks. They shed light on the potential and limitations of these models while providing valuable insights for performance improvement.

4 WHERE TO EVALUATE: DATASETS AND BENCHMARKS

LLMs evaluation datasets are used to test and compare the performance of different language models on various tasks, as depicted in Section 3. These datasets, such as GLUE [200] and SuperGLUE [199], aim to simulate real-world language processing scenarios and cover diverse tasks such as text classification, machine translation, reading comprehension, and dialogue generation. This section will not discuss any single dataset for language models but benchmarks for LLMs.

A variety of benchmarks have emerged to evaluate their performance. In this study, we compile a selection of 46 popular benchmarks, as shown in Table 7.[5] Each benchmark focuses on different aspects and evaluation criteria, providing valuable contributions to their respective domains. For a better summarization, we divide these benchmarks into three categories: benchmarks for general language tasks, benchmarks for specific downstream tasks, and benchmarks for multi-modal tasks.

4.1 Benchmarks for General Tasks

LLMs are designed to solve a vast majority of tasks. To this end, existing benchmarks tend to evaluate the performance in different tasks.

Chatbot Arena [128] and MT-Bench [258] are two significant benchmarks that contribute to the evaluation and advancement of chatbot models and LLMs in different contexts. Chatbot Arena provides a platform to assess and compare diverse chatbot models through user engagement and voting. Users can engage with anonymous models and express their preferences via voting. The platform gathers a significant volume of votes, facilitating the evaluation of models' performance in realistic scenarios. Chatbot Arena provides valuable insights into the strengths and limitations of chatbot models, thereby contributing to the progress of chatbot research and advancement.

Meanwhile, MT-Bench evaluates LLMs on multi-turn dialogues using comprehensive questions tailored to handling conversations. It provides a comprehensive set of questions specifically designed for assessing the capabilities of models in handling multi-turn dialogues. MT-Bench possesses several distinguishing features that differentiate it from conventional evaluation methodologies. Notably, it excels in simulating dialogue scenarios representative of real-world settings, thereby facilitating a more precise evaluation of a model's practical performance. Moreover, MT-Bench effectively overcomes the limitations in traditional evaluation approaches, particularly in gauging a model's competence in handling intricate multi-turn dialogue inquiries.

Instead of focusing on specific tasks and evaluation metrics, HELM [114] provides a comprehensive assessment of LLMs. It evaluates language models across various aspects such as language

[5]Note that as the evaluation of LLMs is a hot research area, it is very likely that we cannot cover all benchmarks. We welcome suggestions and comments to make this list perfect.

Table 7. Summary of Existing LLMs Evaluation Benchmarks (Ordered by the Name of the First Author)

Benchmark	Focus	Domain	Evaluation Criteria
SOCKET [23]	Social knowledge	Specific downstream task	Social language understanding
MME [46]	Multimodal LLMs	Multi-modal task	Ability of perception and cognition
Xiezhi [59]	Comprehensive domain knowledge	General language task	Overall performance across multiple benchmarks
Choice-75 [75]	Script learning	Specific downstream task	Overall performance of LLMs
CUAD [71]	Legal contract review	Specific downstream task	Legal contract understanding
TRUSTGPT [79]	Ethics	Specific downstream task	Toxicity, bias, and value-alignment
MMLU [70]	Text models	General language task	Multitask accuracy
MATH [72]	Mathematical problem	Specific downstream task	Mathematical ability
APPS [68]	Coding challenge competence	Specific downstream task	Code generation ability
CELLO [66]	Complex instructions	Specific downstream task	Four designated evaluation criteria
C-Eval [78]	Chinese evaluation	General language task	52 Exams in a Chinese context
EmotionBench [76]	Empathy ability	Specific downstream task	Emotional changes
OpenLLM [80]	Chatbots	General language task	Leaderboard rankings
DynaBench [94]	Dynamic evaluation	General language task	NLI, QA, sentiment, and hate speech
Chatbot Arena [128]	Chat assistants	General language task	Crowdsourcing and Elo rating system
AlpacaEval [112]	Automated evaluation	General language task	Metrics, robustness, and diversity
CMMLU [108]	Chinese multi-tasking	Specific downstream task	Multi-task language understanding capabilities
HELM [114]	Holistic evaluation	General language task	Multi-metric
API-Bank [109]	Tool utilization	Specific downstream task	API call, retrieval, and planning
M3KE [122]	Multi-task	Specific downstream task	Multi-task accuracy
MMBench [126]	Large vision-language models (LVLMs)	Multi-modal task	Multifaceted capabilities of VLMs
SEED-Bench [107]	Multimodal Large Language Models	Multi-modal task	Generative understanding of MLLMs
UHGEval [116]	Hallucination of Chinese LLMs	Specific downstream task	Form, metric and granularity
ARB [171]	Advanced reasoning ability	Specific downstream task	Multidomain advanced reasoning ability
BIG-bench [182]	Capabilities and limitations of LMs	General language task	Model performance and calibration
MultiMedQA [177]	Medical QA	Specific downstream task	Accuracy and human evaluation
CVALUES [229]	Safety and responsibility	Specific downstream task	Alignment ability of LLMs
LVLM-eHub [230]	LVLMs	Multi-modal task	Multimodal capabilities of LVLMs
ToolBench [191]	Software tools	Specific downstream task	Execution success rate
FRESHQA [198]	Dynamic QA	Specific downstream task	Correctness and hallucination
CMB [211]	Chinese comprehensive medicine	Specific downstream task	Expert evaluation and automatic evaluation
PandaLM [216]	Instruction tuning	General language task	Winrate judged by PandaLM
MINT [213]	Multi-turn interaction	Specific downstream task	Success rate with k-turn budget SR_k
Dialogue CoT [205]	In-depth dialogue	Specific downstream task	Helpfulness and acceptness of LLMs
BOSS [238]	OOD robustness in NLP	General language task	OOD robustness
MM-Vet [237]	Complicated multi-modal tasks	Multi-modal task	Integrated vision-language capabilities
LAMM [234]	Multi-modal point clouds	Multi-modal task	Task-specific metrics
GLUE-X [233]	OOD robustness for NLP tasks	General language task	OOD robustness
KoLA [235]	Knowledge-oriented evaluation	General language task	Self-contrast metrics
AGIEval [260]	Human-centered foundational models	General language task	General
PromptBench [262]	Adversarial prompt resilience	General language task	Adversarial robustness
MT-Bench [258]	Multi-turn conversation	General language task	Winrate judged by GPT-4
M3Exam [248]	Multilingual, multimodal and multilevel	Specific downstream task	Task-specific metrics
GAOKAO-Bench [243]	Chinese Gaokao examination	Specific downstream task	Accuracy and scoring rate
SafetyBench [252]	Safety	Specific downstream task	Safety abilities of LLMs
LLMEval^2 [250]	LLM Evaluator	General language task	Acc, macro-f1 and kappa correlation coefficient

understanding, generation, coherence, context sensitivity, common-sense reasoning, and domain-specific knowledge. HELM aims to holistically evaluate the performance of language models across different tasks and domains. For LLMs Evaluator, Zhang et al. [250] introduces LLMEval^2, which encompasses a wide range of capability evaluations. In addition, Xiezhi [59] presents a comprehensive suite for assessing the knowledge level of large-scale language models in different subject areas. The evaluation conducted through Xiezhi enables researchers to comprehend the notable limitations inherent in these models and facilitates a deeper comprehension of their capabilities in diverse fields. For evaluating language models beyond their existing capacities, BIG-bench [182] introduces a diverse collection of 204 challenging tasks contributed by 450 authors from 132 institutions. These tasks cover various domains such as math, childhood development, linguistics, biology, common-sense reasoning, social bias, physics, software development, and so on.

Recent work has led to the development of benchmarks for evaluating language models' knowledge and reasoning abilities. The **Knowledge-Oriented Language Model Evaluation KoLA** [235] focuses on assessing language models' comprehension and utilization of semantic

knowledge for inference. As such, KoLA serves as an important benchmark for evaluating the depth of language understanding and reasoning in language models, thereby driving progress in language comprehension. To enable crowd-sourced evaluations of language tasks, DynaBench [94] supports dynamic benchmark testing. DynaBench explores new research directions including the effects of closed-loop integration, distributional shift characteristics, annotator efficiency, influence of expert annotators, and model robustness to adversarial attacks in interactive settings. Furthermore, to evaluate language models' ability to learn and apply multidisciplinary knowledge across educational levels, the **Multidisciplinary Knowledge Evaluation M3KE** [122] was recently introduced. M3KE assesses knowledge application within the Chinese education system.

The development of standardized benchmarks for evaluating LLMs on diverse tasks has been an important research focus. MMLU [70] provides a comprehensive suite of tests for assessing text models in multi-task contexts. AlpacaEval [112] stands as an automated evaluation benchmark, which places its focus on assessing the performance of LLMs across various natural language processing tasks. It provides a range of metrics, robustness measures, and diversity evaluations to gauge the capabilities of LLMs. AlpacaEval has significantly contributed to advancing LLMs in diverse domains and promoting a deeper understanding of their performance. Furthermore, AGIEval [260], serves as a dedicated evaluation framework for assessing the performance of foundation models in the domain of human-centric standardized exams. Moreover, OpenLLM [80] functions as an evaluation benchmark by offering a public competition platform for comparing and assessing different LLM models' performance on various tasks. It encourages researchers to submit their models and compete on different tasks, driving progress and competition in LLM research.

As for tasks beyond standard performance, there are benchmarks designed for OOD, adversarial robustness, and fine-tuning. GLUE-X [233] is a novel attempt to create a unified benchmark aimed at evaluating the robustness of NLP models in OOD scenarios. This benchmark emphasizes the significance of robustness in NLP and provides insights into measuring and enhancing the robustness of models. In addition, Yuan et al. [238] presents BOSS, a benchmark collection for assessing out-of-distribution robustness in natural language processing tasks. PromptBench [262] centers on the importance of prompt engineering in fine-tuning LLMs. It provides a standardized evaluation framework to compare different prompt engineering techniques and assess their impact on model performance. PromptBench facilitates the enhancement and optimization of fine-tuning methods for LLMs. To ensure impartial and equitable evaluation, PandaLM [216] is introduced as a discriminative large-scale language model specifically designed to differentiate among multiple high-proficiency LLMs through training. In contrast to conventional evaluation datasets that predominantly emphasize objective correctness, PandaLM incorporates crucial subjective elements, including relative conciseness, clarity, adherence to instructions, comprehensiveness, and formality.

4.2 Benchmarks for Specific Downstream Tasks

Other than benchmarks for general tasks, there exist benchmarks specifically designed for certain downstream tasks.

Question-answering benchmarks have become a fundamental component in the assessment of LLMs and their overall performance. MultiMedQA [177] is a medical QA benchmark that focuses on medical examinations, medical research, and consumer healthcare questions. It consists of seven datasets related to medical QA, including six existing datasets and one new dataset. The goal of this benchmark is to evaluate the performance of LLMs in terms of clinical knowledge and QA abilities. To assess the ability of LLMs in dynamic QA about current world knowledge, Vu et al. [198] introduced FRESHQA. By incorporating relevant and current information retrieved from search engines into prompts, there is a significant enhancement in the performance of LLMs on

FRESHQA. To effectively assess in-depth dialogue, Wang et al. [205] introduced the Dialogue CoT, incorporating two efficient dialogue strategies: Explicit CoT and CoT.

The assessment of LLMs in diverse and demanding tasks has garnered substantial attention in recent research. To this end, a range of specialized benchmarks have been introduced to evaluate LLMs' capabilities in specific domains and applications. Among these, ARB, as presented by Sawada et al. [171], focuses on probing the performance of LLMs in advanced reasoning tasks spanning multiple domains. Additionally, ethical considerations in LLMs have become an area of paramount importance. TRUSTGPT, as tailored by Huang et al. [79], addresses critical ethical dimensions, including toxicity, bias, and value alignment, within the context of LLMs. Furthermore, the simulation of human emotional reactions by LLMs remains an area with significant potential for improvement, as highlighted by the EmotionBench benchmark by Huang et al. [76]. In terms of security evaluation, Zhang et al. [252] have introduced SafetyBench, a benchmark specifically designed to test the security performance of a range of popular Chinese and English LLMs. The results of this evaluation reveal substantial security flaws in current LLMs. To evaluate the daily decision-making capabilities of intelligent systems, Hou et al. [75] introduced Choice-75. Additionally, to assess LLMs' aptitude in understanding complex instructions, He et al. [66] introduced CELLO. This benchmark encompasses the design of eight distinctive features, the development of a comprehensive evaluation dataset, and the establishment of four evaluation criteria alongside their respective measurement standards.

There are other specific benchmarks such as C-Eval [78], which is the first extensive benchmark to assess the advanced knowledge and reasoning capabilities of foundation models in Chinese. Additionally, Li et al. [108] introduces CMMLU as a comprehensive Chinese proficiency standard and evaluates the performance of 18 LLMs across various academic disciplines. The findings reveal that the majority of LLMs demonstrate suboptimal performance in Chinese language environments, highlighting areas for improvement. M3Exam [248] provides a unique and comprehensive evaluation framework that incorporates multiple languages, modalities, and levels to test the general capabilities of LLMs in diverse contexts. Additionally, GAOKAO-Bench [243] provides a comprehensive evaluation benchmark for gauging the proficiency of large language models in intricate and context-specific tasks, utilizing questions sourced from the Chinese Gaokao examination. On the other hand, SOCKET [23] serves as an NLP benchmark designed to evaluate the performance of LLMs in learning and recognizing social knowledge concepts. It consists of several tasks and case studies to assess the limitations of LLMs in social capabilities. MATH [72] concentrates on assessing reasoning and problem-solving proficiencies of AI models within the domain of mathematics. APPS [68] is a more comprehensive and rigorous benchmark for evaluating code generation, measuring the ability of language models to generate Python code according to natural language specifications. CUAD [71] is an expert-annotated, domain-specific legal contract review dataset that presents a challenging research benchmark and potential for enhancing deep learning models' performance in contract understanding tasks. CVALUES [229] introduces a humanistic evaluation benchmark to assess the alignment of LLMs with safety and responsibility standards. In the realm of comprehensive Chinese medicine, Wang et al. [211] introduced CMB, a medical evaluation benchmark rooted in the Chinese language and culture. It addresses the potential inconsistency in the local context that may arise from relying solely on English-based medical assessments. In the realm of hallucination assessment, [116] has developed UHGEval, a benchmark specifically designed to evaluate the performance of Chinese LLMs in text generation without being constrained by hallucination-related limitations.

In addition to existing evaluation benchmarks, there is a research gap in assessing the effectiveness of utilizing tools for LLMs. To address this gap, the API-Bank benchmark [109] is introduced as the first benchmark explicitly designed for tool-augmented LLMs. It comprises a comprehensive

Table 8. Summary of New LLMs Evaluation Protocols

Method	References
Human-in-the-loop	AdaVision [50], AdaTest [164]
Crowd-sourcing testing	DynaBench [94], DynaBoard [132], DynamicTempLAMA [135], DynaTask [188]
More challenging tests	HELM [114], AdaFilter [157], CheckList [165], Big-Bench [182], DeepTest [190]

Tool-Augmented LLM workflow, encompassing 53 commonly used API tools and 264 annotated dialogues, encompassing a total of 568 API calls. Furthermore, the ToolBench project [191] aims to empower the development of large language models that effectively leverage the capabilities of general-purpose tools. By providing a platform for creating optimized instruction datasets, the ToolBench project seeks to drive progress in language models and enhance their practical applications. To evaluate LLMs in multi-turn interactions, Wang et al. [213] proposed MINT, which utilizes tools and natural language feedback.

4.3 Benchmarks for Multi-modal Task

For the evaluation of **Multimodal Large Language Models (MLLMs)**, MME [46] serves as an extensive evaluative benchmark, aiming to assess their perceptual and cognitive aptitudes. It employs meticulously crafted instruction-answer pairs alongside succinct instruction design, thereby guaranteeing equitable evaluation conditions. To robustly evaluate large-scale vision-language models, Liu et al. [126] introduced MMBench, which comprises a comprehensive dataset and employs a CircularEval assessment method. Additionally, MMICL [253] enhances visual language models for multimodal inputs and excels in tasks such as MME and MMBench. Furthermore, LAMM [234] extends its research to encompass multimodal point clouds. LVLM-eHub [230] undertakes an exhaustive evaluation of LVLMs using an online competitive platform and quantitative capacity assessments. To comprehensively assess the generative and understanding capabilities of Multi-modal Large Language Models (MLLMs), Li et al. [107] introduced a novel benchmark named SEED-Bench. This benchmark consists of 19,000 multiple-choice questions that have been annotated by human assessors. Additionally, the evaluation covers 12 different aspects, including the models' proficiency in understanding patterns within images and videos. In summary, recent works have developed robust benchmarks and improved models that advance the study of multimodal languages.

5 HOW TO EVALUATE

In this section, we introduce two common evaluation methods: automatic evaluation and human evaluation. Our categorization is based on whether or not the evaluation criterion can be automatically computed. If it can be automatically calculated, we categorize it into *automatic* evaluation; otherwise, it falls into *human* evaluation.

5.1 Automatic Evaluation

Automated evaluation is a common, and perhaps the most popular, evaluation method that typically uses standard metrics and evaluation tools to evaluate model performance. Compared with human evaluation, automatic evaluation does not require intensive human participation, which not only saves time, but also reduces the impact of human subjective factors and makes the evaluation process more standardized. For example, both Qin et al. [159] and Bang et al. [6] use automated evaluation methods to evaluate a large number of tasks. Recently, with the development of LLMs, some advanced automatic evaluation techniques are also designed to help evaluate. Lin and Chen [121] proposed LLM-EVAL, a unified multidimensional automatic evaluation method for

Table 9. Key Metrics of Automatic Evaluation

General metrics	Metrics
Accuracy	Exact match, Quasi-exact match, F1 score, ROUGE score [118]
Calibrations	Expected calibration error [60], Area under the curve [54]
Fairness	Demographic parity difference [241], Equalized odds difference [64]
Robustness	Attack success rate [203], Performance drop rate [262]

open-domain conversations with LLMs. PandaLM [216] can achieve reproducible and automated language model assessment by training an LLM that serves as the "judge" to evaluate different models. Proposing a self-supervised evaluation framework, Jain et al. [82] enabled a more efficient form of evaluating models in real-world deployment by eliminating the need for laborious labeling of new data. In addition, many benchmarks also apply automatic evaluation, such as MMLU [70], HELM [114], C-Eval [78], AGIEval [260], AlpacaFarm [38], Chatbot Arena [128], and the like.

Based on the literature that adopted automatic evaluation, we summarized the main metrics in automatic evaluation in Table 9. The key metrics include the following four aspects:

(1) **Accuracy** is a measure of how correct a model is on a given task. The concept of accuracy may vary in different scenarios and is dependent on the specific task and problem definition. It can be measured using various metrics such as Exact Match, F1 score, and ROUGE score.
   — **Exact Match (EM)** is a metric used to evaluate whether the model's output in text generation tasks precisely matches the reference answer. In question answering tasks, if the model's generated answer is an exact match with the manually provided answer, the EM is 1; otherwise, it is 0.
   — The F1 score is a metric for evaluating the performance of binary classification models, combining the model's precision and recall. The formula for calculation is as follows: $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$.
   — ROUGE is primarily employed to assess the performance of tasks such as text summarization and machine translation, involving considerations of overlap and matching between texts.

(2) **Calibrations** pertains to the degree of agreement between the confidence level of the model output and the actual prediction accuracy.
   — **Expected Calibration Error (ECE)** is one of the commonly used metrics to evaluate model calibration performance [60]. Tian et al. [189] utilized ECE to study the calibration of RLHF-LMs, including ChatGPT, GPT-4, Claude 1, Claude 2 and Llama2. For the calculation of ECE, they categorize model predictions based on confidence and measure the average accuracy of the predictions within each confidence interval.
   — **Area Under the Curve** of selective accuracy and coverage **(AUC)** [54] is another commonly used metric.

(3) **Fairness** refers to whether the model treats different groups consistently, that is, whether the model's performance is equal across different groups. This can include attributes such as gender, race, age, and more. DecodingTrust [201] employs the following two metrics for measuring fairness:
   — **Demographic Parity Pifference (DPD)** measures whether the model's predictions are distributed equally across different population groups. If predictions differ significantly between groups, the DPD is high, indicating that the model may be unfairly biased against different groups. The calculation of DPD involves the prediction of the model and the true label, and the following formula can be used: $DPD = P(\hat{y}|Z = 1) - P(\hat{y}|Z = 0)$, where

$\hat{y}$ is the binary classification prediction of the model, Z is the identifier of the population group (usually binary, indicating two different groups, such as men and women), $P(\hat{y}|Z = 1)$ and $P(\hat{y}|Z = 0)$ respectively represent the probabilities of predicting the positive class in population $Z = 1$ and $Z = 0$.

— **Equalized Odds Difference (EOD)** aims to ensure that the model provides equal error rates across different populations, that is, the model's prediction error probability distribution is similar for different populations. The calculation of EOD involves probabilities related to **true positive (TP)**, **true negative (TN)**, **false positive (FP)**, and **false negative (FN)** predictions. The formula for EOD is as follows: $max\{P(\hat{y} = 1|Y = 1, Z = 1) - P(\hat{y} = 1|Y = 1, Z = 0), P(\hat{y} = 1|Y = 0, Z = 1) - P(\hat{y} = 1|Y = 0, Z = 0)\}$ where $\hat{y}$ is the binary classification prediction of the model, Y is the true label, Z is the demographic group identifier (typically binary, representing two different groups), and $P(\hat{y} = 1|Y = 1, Z = 1)$ denotes the probability of the model predicting a positive class when the true label is positive and belongs to group $Z = 1$.

(4) **Robustness** evaluates the performance of a model in the face of various challenging inputs, including adversarial attacks, changes in data distribution, noise, and so o.

— **Attack Success Rate (ASR)** serves as a metric for evaluating the adversarial robustness of LLMs [206]. Specifically, consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ containing N pairs of samples x_i and ground truth y_i. For an adversarial attack method $\mathcal{A}$, given an input x, this method can produce adversarial examples $\mathcal{A}(x)$ to attack surrogate model f, with the success rate is calculated as: $ASR = \sum_{(x,y\in D)} \frac{\mathcal{I}[f(\mathcal{A}(x))\neq y]}{\mathcal{I}[f(x)=y]}$, where $\mathcal{I}$ is the indicator function [203].

— **Performance Drop Rate (PDR)**, a new unified metric, effectively assesses the robustness of prompt in LLMs [262]. PDR quantifies the relative performance degradation after a prompt attack, and the formula is as follows: $PDR = 1 - \frac{\sum_{(x,y)\in D} \mathcal{M}[f([A(P),x]),y]}{\sum_{(x,y)\in D} \mathcal{M}[f([P,x]),y]}$, where A represents the adversarial attack applied to prompt P, and M denotes the evaluation function, which varies across different tasks [262].

5.2 Human Evaluation

The increasingly strengthened capabilities of LLMs have certainly gone beyond standard evaluation metrics on general natural language tasks. Therefore, human evaluation becomes a natural choice in some non-standard cases where automatic evaluation is not suitable. For instance, in open-generation tasks where embedded similarity metrics (such as BERTScore) are not enough, human evaluation is more reliable [142]. While some generation tasks can adopt certain automatic evaluation protocols, human evaluation in these tasks is more favorable as generation can always go better than standard answers.

Human evaluation is a way to evaluate the quality and accuracy of model-generated results through human participation. Compared with automatic evaluation, manual evaluation is closer to the actual application scenario and can provide more comprehensive and accurate feedback. In the manual evaluation of LLMs, evaluators (such as experts, researchers, or ordinary users) are usually invited to evaluate the results generated by the model. For example, Ziems et al. [267] used the annotations from experts for generation. By human evaluation, Liang et al. [114] assessed on summarization and disinformation scenarios on six models and Bang et al. [6] evaluated analogical reasoning tasks. Bubeck et al. [15] did a series of human-crafted tests using GPT-4 and they found that GPT-4 performs close to or even exceeds human performance on multiple tasks. This evaluation requires human evaluators to actually test and compare the performance of the models, not just evaluate the models through automated evaluation metrics. Note that even human evaluations can have high variance and instability, which could be due to cultural and

Table 10. Summary of Key Factors in Human Evaluation

Evaluation Criteria	Key Factor
Number of evaluators	Adequate representation [7], Statistical significance
Evaluation rubrics	Accuracy [178], Relevance [259], Fluency [196], Transparency, Safety [85], Human alignment
Evaluator's expertise level	Relevant domain expertise [144], Task familiarity, Methodological training

individual differences [155]. In practical applications, these two evaluation methods are considered and weighed in combination with the actual situation.

Exploring the human evaluation methods of LLMs requires thoughtful attention to various crucial factors to guarantee the dependability and precision of assessments [178]. Table 10 provides a concise overview of the essential aspects of human evaluation, including the number of evaluators, evaluation criteria, and evaluator's expertise level. Primarily, the number of evaluators emerges as a crucial factor intricately intertwined with adequate representation and statistical significance. A judiciously chosen number of evaluators contributes to a more nuanced and comprehensive understanding of the LLMs under scrutiny, enabling a more reliable extrapolation of the results to a broader context.

Furthermore, evaluation criteria are fundamental components of the human assessment process. Expanding upon the principles of the **3H rule (Helpfulness, Honesty, and Harmlessness)** [4], we have elaborated them into the following six human assessment criteria. These criteria include accuracy, relevance, fluency, transparency, safety, and human alignment. Through the application of these standards, a thorough analysis of LLMs' performance in syntax, semantics, and context is achieved, allowing for a more comprehensive evaluation of the quality of generated text.

(1) **Accuracy** [178] stands out as a pivotal criterion that assesses the precision and correctness of the generated text. It involves scrutinizing the extent to which the language model produces information that aligns with factual knowledge, avoiding errors and inaccuracies.
(2) **Relevance** [259] focuses on the appropriateness and significance of the generated content. This criterion examines how well the text addresses the given context or query, ensuring that the information provided is pertinent and directly applicable.
(3) **Fluency** [196] assesses the language model's ability to produce content that flows smoothly, maintaining a consistent tone and style. A fluent text is not only grammatically correct but also ensures readability and a seamless user experience. Analysts evaluate how well the model avoids awkward expressions and abrupt shifts in language or topic, contributing to effective communication with users.
(4) **Transparency** delves into the clarity and openness of the language model's decision-making process. It involves assessing how well the model communicates its thought processes, enabling users to understand how and why certain responses are generated. A transparent model provides insights into its inner workings.
(5) **Safety** [85] emerges as a critical criterion concerned with the potential harm or unintended consequences arising from the generated text. It examines the language model's ability to avoid producing content that may be inappropriate, offensive, or harmful, ensuring the well-being of users and avoiding misinformation.
(6) **Human alignment** assesses the degree to which the language model's output aligns with human values, preferences, and expectations. It considers the ethical implications of the generated content, ensuring that the language model produces text that respects societal norms and user expectations, promoting a positive interaction with human users.

Lastly, the expertise level of evaluators is a critical consideration, encompassing relevant domain knowledge, task familiarity, and methodological training. Delineating the requisite expertise level for evaluators ensures that they possess the necessary background knowledge to accurately comprehend and assess the domain-specific text generated by LLMs. This strategy adds a layer of rigor to the evaluation process, reinforcing the credibility and validity of the findings.

6 SUMMARY

In this section, we summarize the key findings based on our review in Sections 3, 4, and 5.

First of all, we would like to highlight that despite all the efforts spent on summarizing existing works on evaluation, there is *no* evidence to explicitly show that one certain evaluation protocol or benchmark is the most useful and successful, but with **different characteristics and focuses**. This also demonstrates that not a single model can perform best in all kinds of tasks. The purpose of this survey is to go beyond simply determining the "best" benchmark or evaluation protocol. By summarizing and analyzing existing efforts on LLMs evaluation, we may identify the current success and failure cases of LLMs, derive new trends for evaluation protocols, and most importantly, propose new challenges and opportunities for future research.

6.1 Task: Success and Failure Cases of LLMs

We now summarize the success and failure cases of LLMs in different tasks. Note that all the following conclusions are made based on existing evaluation efforts and the results are only dependent on specific datasets.

6.1.1 What Can LLMs do Well?

- LLMs demonstrate proficiency in generating text [11, 14, 24] by producing fluent and precise linguistic expressions.
- LLMs obtain impressive performance in tasks involving language understanding, including sentiment analysis [52, 129, 159], text classification [114, 154, 232], as well as the handling of factual input [159].
- LLMs demonstrate robust arithmetic reasoning capabilities [159] and excel in logical reasoning [124]. Moreover, they exhibit noteworthy proficiency in temporal reasoning [6]. Furthermore, more intricate tasks such as mathematical reasoning [225, 236, 243] and structured data inference [86, 151] have emerged as the prevailing benchmarks for evaluation.
- LLMs exhibit robust contextual comprehension, enabling them to generate coherent responses that align with the given input [187].
- LLMs also achieve satisfying performance across several natural language processing tasks, including machine translation [6, 130, 208], text generation [20], and question answering [102, 114].

6.1.2 When Can LLMs Fail?

- Within the realm of NLI, LLMs exhibit subpar performance and encounter challenges in accurately representing human disagreements [105].
- LLMs exhibit restricted proficiency in discerning semantic similarity between events [184] and demonstrate substandard performance in evaluating fundamental phrases [166].
- LLMs have limited abilities on abstract reasoning [56], and are prone to confusion or errors in complex contexts [148].
- In linguistic contexts featuring non-Latin scripts and limited resources, LLMs manifest suboptimal performance [2, 6, 100, 248]. Furthermore, generative LLMs consistently display proficiency levels below the expected standards across various tasks and languages [2].

— LLMs demonstrate susceptibility when processing visual modal information [256]. Furthermore, they have the capacity to assimilate, disseminate, and potentially magnify detrimental content found within the acquired training datasets, frequently encompassing toxic linguistic elements, including offensive, hostile, and derogatory language [53].
— LLMs may exhibit social biases and toxicity [37, 53, 153] during the generation process, resulting in the production of biased outputs.
— LLMs may manifest credibility deficits [201], potentially giving rise to fabricated information or erroneous facts within dialogues [163, 251].
— LLMs have limitations in incorporating real-time or dynamic information [127], making them less suitable for tasks that require up-to-date knowledge or rapid adaptation to changing contexts.
— LLMs are sensitive to prompts, especially adversarial prompts [262], which trigger new evaluations and algorithms to improve its robustness.

6.2 Benchmark and Evaluation Protocol

With the rapid development and widespread use of LLMs, the importance of evaluating them in practical applications and research has become crucial. This evaluation process should include not only task-level evaluation but also a deep understanding of the potential risks they pose from a societal perspective. In this section, we summarize existing benchmarks and protocols in Table 8.

First, a shift from objective calculation to human-in-the-loop testing, allowing for greater human feedback during the evaluation process. AdaVision [50], an interactive process for testing vision models, enables users to label a small amount of data for model correctness, which helps users identify and fix coherent failure modes. In AdaTest [164], the user filters test samples by only selecting high-quality tests and organizing them into semantically related topics.

Second, a move from static to crowd-sourcing test sets is becoming more common. Tools like DynaBench [94], DynaBoard [132], and DynaTask [188] rely on crowdworkers to create and test hard samples. Additionally, DynamicTempLAMA [135] allows for dynamically constructed time-related tests.

Third, a shift from a unified to a challenging setting in evaluating machine learning models. While unified settings involve a test set with no preference for any specific task, challenging settings create test sets for specific tasks. Tools like DeepTest [190] use seeds to generate input transformations for testing, CheckList [165] builds test sets based on templates, and AdaFilter [157] adversarially constructs tests. However, it is worth noting that AdaFilter may not be entirely fair as it relies on adversarial examples. HELM [114] evaluates LLMs from different aspects, while the Big-Bench [182] platform is used to design hard tasks for machine learning models to tackle. PromptBench [262] aims to evaluate the adversarial robustness of LLMs by creating adversarial prompts, which is more challenging and the results demonstrated that current LLMs are not robust to adversarial prompts.

7 GRAND CHALLENGES AND OPPORTUNITIES FOR FUTURE RESEARCH

Evaluation as a new discipline: Our summarization inspires us to redesign a wide spectrum of aspects related to evaluation in the era of LLMs. In this section, we present several grand challenges. Our key point is that **evaluation should be treated as an essential discipline to drive the success of LLMs and other AI models**. Existing protocols are not enough to thoroughly evaluate the true capabilities of LLMs, which poses grand challenges and triggers new opportunities for future research on LLMs evaluation.

7.1 Designing AGI Benchmarks

As we discussed earlier, while all tasks can potentially serve as evaluation tools for LLMs, the question remains as to which can truly measure AGI capabilities. As we expect LLMs to demonstrate AGI abilities, a comprehensive understanding of the differences between human and AGI capacities becomes crucial in the creation of AGI benchmarks. The prevailing trend seems to conceptualize AGI as a superhuman entity, thereby utilizing cross-disciplinary knowledge from fields such as education, psychology, and social sciences to design innovative benchmarks. Nonetheless, there remains a plethora of unresolved issues. For instance, does it make sense to use human values as a starting point for test construction, or should alternative perspectives be considered? Developing suitable AGI benchmarks presents many open questions demanding further exploration.

7.2 Complete Behavioral Evaluation

An ideal AGI evaluation should contain not only standard benchmarks on common tasks, but also evaluations on open tasks such as complete behavioral tests. By behavioral test, we mean that AGI models should also be evaluated in an open environment. For instance, by treating LLMs as the central controller, we can construct evaluations on a robot manipulated by LLMs to test its behaviors in real situations. By treating LLMs as a completely intelligent machine, the evaluations of its multi-modal dimensions should also be considered. In fact, complete behavioral evaluations are complementary to standard AGI benchmarks and they should work together for better testing.

7.3 Robustness Evaluation

Beyond general tasks, it is crucial for LLMs to maintain robustness against a wide variety of inputs in order to perform optimally for end-users, given their extensive integration into daily life. For instance, the same prompts but with different grammars and expressions could lead ChatGPT and other LLMs to generate diverse results, indicating that current LLMs are not robust to the inputs. While there are some prior works on robustness evaluation [206, 262], there is much room for advancement, such as including more diverse evaluation sets, examining more evaluation aspects, and developing more efficient evaluations to generate robustness tasks. Concurrently, the concept and definition of robustness are constantly evolving. It is thus vital to consider updating the evaluation system to better align with emerging requirements related to ethics and bias.

7.4 Dynamic and Evolving Evaluation

Existing evaluation protocols for most AI tasks rely on static and public benchmarks, i.e., the evaluation datasets and protocols are often publicly available. While this facilitates rapid and convenient evaluation within the community, it is unable to accurately assess the evolving abilities of LLMs, given their rapid rate of development. The capabilities of LLMs may enhance over time which cannot be consistently evaluated by existing static benchmarks. On the other hand, as LLMs grow increasingly powerful with larger model sizes and training set sizes, static and public benchmarks are likely to be memorized by LLMs, resulting in potential training data contamination. Therefore, developing dynamic and evolving evaluation systems is the key to providing a fair evaluation of LLMs.

7.5 Principled and Trustworthy Evaluation

When introducing an evaluation system, it is crucial to ascertain its integrity and trustworthiness. Therefore, the necessity for trustworthy computing extends to the requirement for reliable evaluation systems as well. This poses a challenging research question that intertwines with

measurement theory, probability, and numerous other domains. For instance, how can we ensure that dynamic testing truly generates out-of-distribution examples? There is a scarcity of research in this domain, and it is hoped that future work will aim to scrutinize not only the algorithms but the evaluation system itself.

7.6 Unified Evaluation that Supports All LLMs Tasks

There are many other research areas of LLMs and we need to develop evaluation systems that can support all kinds of tasks such as value alignment, safety, verification, interdisciplinary research, fine-tuning, and others. For instance, PandaLM [216] is an evaluation system that assists LLMs fine-tuning by providing an open-source evaluation model, which can automatically assess the performance of fine-tuning. We expect that more evaluation systems are becoming more general and can be used as assistance in certain LLMs tasks.

7.7 Beyond Evaluation: LLMs Enhancement

Ultimately, evaluation is not the end goal but rather the starting point. Following the evaluation, there are undoubtedly conclusions to be drawn regarding performance, robustness, stability, and other factors. A proficient evaluation system should not only offer benchmark results but should also deliver an insightful analysis, recommendations, and guidance for future research and development. For instance, PromptBench [262] provides not only robustness evaluation results on adversarial prompts but also a comprehensive analysis through attention visualization, elucidating how adversarial texts can result in erroneous responses. The system further offers a word frequency analysis to identify robust and non-robust words in the test sets, thus providing prompt engineering guidance for end users. Subsequent research can leverage these findings to enhance LLMs. Another example is that Wang et al. [215] first explored the performance of large vision-language models on imbalanced (long-tailed) tasks, which demonstrates the limitation of current large models. Then, they explored different methodologies to enhance the performance on these tasks. In summary, enhancement after evaluation helps to build better LLMs and much can be done in the future.

8 CONCLUSION

Evaluation carries profound significance, becoming imperative in the advancement of AI models, especially within the context of large language models. This paper presents the first survey to give a comprehensive overview of the evaluation on LLMs from three aspects: what to evaluate, how to evaluate, and where to evaluate. By encapsulating evaluation tasks, protocols, and benchmarks, our aim is to augment understanding of the current status of LLMs, elucidate their strengths and limitations, and furnish insights for future LLMs progression.

Our survey reveals that current LLMs exhibit certain limitations in numerous tasks, notably reasoning and robustness tasks. Concurrently, the need for contemporary evaluation systems to adapt and evolve remains evident, ensuring the accurate assessment of LLMs' inherent capabilities and limitations. We identify several grand challenges that future research should address, with the aspiration that LLMs can progressively enhance their service to humanity.

DISCLAIMER

The goal of this paper is mainly to summarize and discuss existing evaluation efforts on large language models. Results and conclusions in each paper are original contributions of their corresponding authors, particularly for potential issues in ethics and biases. This paper may discuss some side effects of LLMs and the only intention is to foster a better understanding.

Due to the evolution of LLMs especially online services such as Claude and ChatGPT, it is very likely that they become stronger and some of the limitations described in this paper are mitigated (and new limitations may arise). We encourage interested readers to take this survey as a reference for future research and conduct real experiments in current systems when performing evaluations.

Finally, the evaluation of LLMs is continuously developing, thus we may miss some new papers or benchmarks. We welcome all constructive feedback and suggestions.

REFERENCES

[1] Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2023. Benchmarking Arabic AI with large language models. *arXiv preprint arXiv:2305.14982* (2023).

[2] Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. MEGA: Multilingual evaluation of generative AI. *arXiv preprint arXiv:2303.12528* (2023).

[3] Daman Arora, Himanshu Gaurav Singh, et al. 2023. Have LLMs advanced enough? A challenging problem solving benchmark for large language models. *arXiv preprint arXiv:2305.15074* (2023).

[4] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).

[5] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181* (2023).

[6] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[7] Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. 313–320.

[8] Daniel Berrar. 2019. Cross-Validation. (2019).

[9] Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421* (2023).

[10] Bojana Bodroza, Bojana M. Dinic, and Ljubisa Bojic. 2023. Personality testing of GPT-3: Limited temporal reliability, but highlighted social desirability of GPT-3's personality instruments results. *arXiv preprint arXiv:2306.04308* (2023).

[11] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[12] Nathan Brody. 1999. What is intelligence? *International Review of Psychiatry* 11, 1 (1999), 19–25.

[13] Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18, 4 (1992), 467–480.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[15] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).

[16] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. 53–67.

[17] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *Journal of Medical Systems* 47, 1 (2023), 33.

[18] Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? A conversation with ChatGPT. *Journal of Chemical Information and Modeling* 63, 6 (2023), 1649–1655.

[19] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[20] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723* (2023).
[21] Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. 2023. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility* (2023).
[22] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757* (2023).
[23] Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? Evaluating the sociability of large language models with SocKET benchmark. *arXiv preprint arXiv:2305.14938* (2023).
[24] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).
[25] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* 30 (2017).
[26] Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. 2023. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 3–17.
[27] Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, et al. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694* (2023).
[28] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (1995), 273–297.
[29] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's capabilities in recommender systems. *arXiv preprint arXiv:2305.02182* (2023).
[30] Wei Dai, Jionghao Lin, Flora Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gasevic, and Guanliang Chen. 2023. Can large language models provide feedback to students? A case study on ChatGPT. (2023).
[31] Xuan-Quy Dao and Ngoc-Bich Le. 2023. Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the Vietnamese national high school graduation examination. *arXiv preprint arXiv:2306.06331* (2023).
[32] Joost C. F. de Winter. 2023. Can ChatGPT pass high school exams on English language comprehension. *Researchgate. Preprint* (2023).
[33] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.
[34] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization? *arXiv preprint arXiv:2306.01248* (2023).
[35] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).
[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[37] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 862–872.
[38] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387* (2023).
[39] Dat Duong and Benjamin D. Solomon. 2023. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics* (2023), 1–3.
[40] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender Systems in the Era of Large Language Models (LLMs). (2023). arXiv:cs.IR/2307.02046
[41] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A new dataset for automatic medical diagnosis. *Advances in Neural Information Processing Systems* 35 (2022), 31306–31318.
[42] Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).
[43] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30 (2020), 681–694.
[44] Michael C. Frank. 2023. Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology* (2023), 1–2.

[45] Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867* (2023).
[46] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394* (2023).
[47] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306* (2023).
[48] Tadayoshi Fushiki. 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing* 21 (2011), 137–146.
[49] Stephen I. Gallant et al. 1990. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks* 1, 2 (1990), 179–191.
[50] Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. 2022. Adaptive testing of computer vision models. *arXiv preprint arXiv:2212.02774* (2022).
[51] Jianfeng Gao and Chin-Yew Lin. 2004. Introduction to the special issue on statistical language modeling. (2004), 87–93.
[52] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).
[53] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3356–3369.
[54] Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems* 30 (2017).
[55] Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171* (2023).
[56] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not abstract reasoners. *arXiv preprint arXiv:2305.19555* (2023).
[57] Aidan Gilson, Conrad W. Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education* 9, 1 (2023), e45312.
[58] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in Experimental Social Psychology*. Vol. 47. Elsevier, 55–130.
[59] Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Qianyu He, Rui Xu, et al. 2023. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. *arXiv preprint arXiv:2306.05783* (2023).
[60] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. PMLR, 1321–1330.
[61] Taicheng Guo, Kehan Guo, Zhengwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What indeed can GPT models do in chemistry? A comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365* (2023).
[62] Thilo Hagendorff and Sarah Fabi. 2023. Human-like Intuitive Behavior and Reasoning Biases Emerged in Language Models – and Disappeared in GPT-4. (2023). arXiv:cs.CL/2306.07622
[63] Alaleh Hamidi and Kirk Roberts. 2023. Evaluation of AI chatbots for patient-specific EHR questions. *arXiv preprint arXiv:2306.02549* (2023).
[64] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016).
[65] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768* (2023).
[66] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, et al. 2023. Can large language models understand real-world complex instructions? *arXiv preprint arXiv:2309.09150* (2023).
[67] Arto Hellas, Juho Leinonen, Sami Sarsa, Charles Koutcheme, Lilja Kujanpää, and Juha Sorva. 2023. Exploring the responses of large language models to beginner programmers' help requests. *arXiv preprint arXiv:2306.05715* (2023).
[68] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938* (2021).

[69] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275* (2020).
[70] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
[71] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An expert-annotated NLP dataset for legal contract review. *arXiv preprint arXiv:2103.06268* (2021).
[72] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).
[73] Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T. Sio, Lisa A. McGee, Jonathan B. Ashman, Xiang Li, Tianming Liu, Jiajian Shen, et al. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *arXiv preprint arXiv:2304.01938* (2023).
[74] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991* (2022).
[75] Zhaoyi Joey Hou, Li Zhang, and Chris Callison-Burch. 2023. Choice-75: A dataset on decision branching in script learning. *arXiv preprint arXiv:2309.11737* (2023).
[76] Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2023. Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. *arXiv preprint arXiv:2308.03656* (2023).
[77] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045* (2023).
[78] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322* (2023).
[79] Yue Huang, Qihui Zhang, Philip S. Y., and Lichao Sun. 2023. TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models. (2023). arXiv:cs.CL/2306.11507
[80] HuggingFace. 2023. Open-source Large Language Models Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard (2023).
[81] Israt Jahan, Md. Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. *arXiv preprint arXiv:2306.04504* (2023).
[82] Neel Jain, Khalid Saifullah, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Bring your own data! Self-supervised evaluation for large language models. *arXiv preprint arXiv:2306.13651* (2023).
[83] Malin Jansson, Stefan Hrastinski, Stefan Stenbom, and Fredrik Enoksson. 2021. Online question and answer sessions: How students support their own and other students' processes of inquiry in a text-based learning environment. *The Internet and Higher Education* 51 (2021), 100817.
[84] Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! Humor is still challenging large language models. *arXiv preprint arXiv:2306.04563* (2023).
[85] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. *arXiv preprint arXiv:2307.04657* (2023).
[86] Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645* (2023).
[87] Douglas Johnson, Rachel Goodman, J. Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. 2023. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. (2023).
[88] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.
[89] Saurav Kadavath, Tom Conerly, Amanda Askell, T. J. Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv* abs/2207.05221 (2022).

[90] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. 2022. MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445* (2022).
[91] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.
[92] Jean Khalfa. 1994. What is intelligence? (1994).
[93] Yousuf A. Khan, Clarisse Hokia, Jennifer Xu, and Ben Ehlert. 2023. covLLM: Large language models for COVID-19 biomedical literature. *arXiv preprint arXiv:2306.04926* (2023).
[94] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).
[95] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, Vol. 14. Montreal, Canada, 1137–1145.
[96] Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukás Burget. 2011. Recurrent neural network based language modeling in meeting recognition. In *Interspeech*, Vol. 11. 2877–2880.
[97] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2, 2 (2023), e0000198.
[98] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics* (2019).
[99] Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S. Glicksberg, and Eyal Klang. 2023. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific Reports* 13, 1 (2023), 4164.
[100] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613* (2023).
[101] Pier Luca Lanzi and Daniele Loiacono. 2023. ChatGPT and other large language models as evolutionary engines for online interactive collaborative game design. *arXiv preprint arXiv:2303.02155* (2023).
[102] Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md. Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv preprint arXiv:2305.18486* (2023).
[103] Van-Hoang Le and Hongyu Zhang. 2023. An evaluation of log parsing with ChatGPT. *arXiv preprint arXiv:2306.01590* (2023).
[104] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[105] Noah Lee, Na Min An, and James Thorne. 2023. Can large language models infer and disagree like humans? *arXiv preprint arXiv:2305.13788* (2023).
[106] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
[107] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).
[108] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. CMMLU: Measuring massive multitask language understanding in Chinese. *arXiv preprint arXiv:2306.09212* (2023).
[109] Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-Bank: A Benchmark for Tool-Augmented LLMs. (2023). arXiv:cs.CL/2304.08244
[110] Ruyu Li, Wenhao Deng, Yu Cheng, Zheng Yuan, Jiaqi Zhang, and Fajie Yuan. 2023. Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights. *arXiv preprint arXiv:2305.11700* (2023).
[111] Xinzhe Li, Ming Liu, Shang Gao, and Wray Buntine. 2023. A Survey on Out-of-Distribution Evaluation of Neural NLP Models. (2023). arXiv:cs.CL/2306.15261
[112] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval (2023).

[113] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).
[114] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).
[115] Tian Liang, Zhiwei He, Jen-tes Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Leveraging word guessing games to assess the intelligence of large language models. *arXiv preprint arXiv:2310.20499* (2023).
[116] Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Zhaohui Wy, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296* (2023).
[117] Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143* (2022).
[118] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013
[119] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).
[120] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.
[121] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711* (2023).
[122] Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023. M3KE: A Massive Multi-Level Multi-Subject Knowledge Evaluation Benchmark for Chinese Large Language Models. (2023). arXiv:cs.CL/2305.10263
[123] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. (2023). arXiv:cs.CV/2306.14565
[124] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4. (2023). arXiv:cs.CL/2304.03439
[125] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210* (2023).
[126] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is Your Multi-modal Model an All-around Player? (2023). arXiv:cs.CV/2307.06281
[127] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852* (2023).
[128] LMSYS. 2023. Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings. https://lmsys.org (2023).
[129] Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619* (2023).
[130] Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with ChatGPT. *arXiv preprint arXiv:2305.01181* (2023).
[131] Qing Lyu, Josh Tan, Mike E. Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T. Whitlow. 2023. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038* (2023).
[132] Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. *Advances in Neural Information Processing Systems* 34 (2021), 10351–10367.
[133] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).
[134] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. MQAG: Multiple-choice Question Answering and Generation for Assessing Information Consistency in Summarization. (2023). arXiv:cs.CL/2301.12307
[135] Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic benchmarking of masked language models on temporal concept drift with multiple views. *arXiv preprint arXiv:2302.12297* (2023).
[136] John McCarthy. 2007. What is artificial intelligence. (2007).
[137] Microsoft. 2023. Bing chat. *https://www.bing.com/new* (2023).

[138] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251* (2023).
[139] John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. *arXiv preprint arXiv:2306.07075* (2023).
[140] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599* (2019).
[141] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. CodeGen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).
[142] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. *arXiv preprint arXiv:1707.06875* (2017).
[143] Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. 2023. ChatGPT goes to the operating room: Evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research* 104, 5 (2023), 269.
[144] Andrew M. Olney. 2023. Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. In *AIED Workshops*.
[145] OpenAI. 2023. https://chat.openai.com.chat (2023).
[146] OpenAI. 2023. GPT-4 Technical Report. (2023). arXiv:cs.CL/2303.08774
[147] Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence* 6 (2023).
[148] Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. ThoughtSource: A central hub for large language model reasoning data. *arXiv preprint arXiv:2301.11596* (2023).
[149] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[150] Vishal Pallagani, Bharath Muppasani, Keerthiram Murugesan, Francesca Rossi, Biplav Srivastava, Lior Horesh, Francesco Fabiano, and Andrea Loreggia. 2023. Understanding the capabilities of large language models for automated planning. *arXiv preprint arXiv:2305.16151* (2023).
[151] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. (2023). arXiv:cs.CL/2306.08302
[152] Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. TALM: Tool augmented language models. *arXiv preprint arXiv:2205.12255* (2022).
[153] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2086–2105.
[154] Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Iñigo Puente, Jorge Cordova, and Gonzalo Cordova. 2023. Leveraging large language models for topic classification in the domain of public affairs. *arXiv preprint arXiv:2306.02864* (2023).
[155] Kaiping Peng, Richard E. Nisbett, and Nancy Y. C. Wong. 1997. Validity problems comparing values across cultures and possible solutions. *Psychological Methods* 2, 4 (1997), 329.
[156] Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. *arXiv preprint arXiv:2306.06264* (2023).
[157] Jason Phang, Angelica Chen, William Huang, and Samuel R. Bowman. 2021. Adversarially constructed evaluation sets are more challenging, but may not be fair. *arXiv preprint arXiv:2111.08181* (2021).
[158] Dongqi Pu and Vera Demberg. 2023. ChatGPT vs Human-authored Text: Insights into Controllable Text Summarization and Sentence Style Transfer. (2023). arXiv:cs.CL/2306.07799
[159] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).
[160] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool Learning with Foundation Models. (2023). arXiv:cs.CL/2304.08354

[161] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. (2023). arXiv:cs.AI/2307.16789

[162] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[163] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).

[164] Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive testing and debugging of NLP models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3253–3267.

[165] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[166] Nicholas Riccardi and Rutvik H. Desai. 2023. The two word test: A semantic benchmark for large language models. *arXiv preprint arXiv:2306.04610* (2023).

[167] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of ChatGPT. *arXiv preprint arXiv:2304.07333* (2023).

[168] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).

[169] Jamil S. Samaan, Yee Hui Yeo, Nithya Rajeev, Lauren Hawley, Stuart Abel, Wee Han Ng, Nitin Srinivasan, Justin Park, Miguel Burch, Rabindra Watson, et al. 2023. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity Surgery* (2023), 1–7.

[170] Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. *arXiv preprint arXiv:2305.15269* (2023).

[171] Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. ARB: Advanced Reasoning Benchmark for Large Language Models. (2023). arXiv:cs.CL/2307.13692

[172] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761* (2023).

[173] Prabin Sharma, Kisan Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. 2023. Performance of ChatGPT on USMLE: Unlocking the potential of large language models for AI-assisted medical education. *arXiv preprint arXiv:2307.00112* (2023).

[174] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace. *arXiv preprint arXiv:2303.17580* (2023).

[175] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4275–4293.

[176] Gabriel Simmons. 2022. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106* (2022).

[177] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138* (2022).

[178] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[179] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990* (2022).

[180] Xiaoyang Song, Akshat Gupta, Kiyan Mohebbizadeh, Shujie Hu, and Anant Singh. 2023. Have large language models developed a personality?: Applicability of self-assessment tests in measuring personality in LLMs. *arXiv preprint arXiv:2305.14693* (2023).

[181] Giriprasad Sridhara, Sourav Mazumdar, et al. 2023. ChatGPT: A study on its utility for ubiquitous software engineering tasks. *arXiv preprint arXiv:2305.16837* (2023).

[182] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md. Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[183] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? Investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542* (2023).
[184] Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Yanlin Feng, Jia Li, and Wenpeng Hu. 2023. EvEval: A comprehensive evaluation of event semantics for large language models. *arXiv preprint arXiv:2305.15268* (2023).
[185] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663* (2021).
[186] Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. 2023. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education* 9, 1 (2023), e46599.
[187] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
[188] Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. 2022. Dynatask: A framework for creating dynamic AI benchmark tasks. *arXiv preprint arXiv:2204.01906* (2022).
[189] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975* (2023).
[190] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*. 303–314.
[191] ToolBench. 2023. Open-source tools learning benchmarks. https://github.com/sambanova/toolbench (2023).
[192] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[193] Alan M. Turing. 2009. *Computing Machinery and Intelligence*. Springer.
[194] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models–a critical investigation. *arXiv preprint arXiv:2305.15771* (2023).
[195] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). *arXiv preprint arXiv:2206.10498* (2022).
[196] Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*. 355–368.
[197] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[198] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. (2023). arXiv:cs.CL/2310.03214
[199] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems* 32 (2019).
[200] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).
[201] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. (2023). arXiv:cs.CL/2306.11698
[202] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. (2021).
[203] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840* (2021).
[204] Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023. Evaluating open question answering evaluation. *arXiv preprint arXiv:2305.12421* (2023).
[205] Hongru Wang, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. 2023. Chain-of-thought prompting for responding to in-depth dialogue questions with LLM. (2023). arXiv:cs.CL/2305.11792

[206] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. In *ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

[207] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[208] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210* (2023).

[209] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926* (2023).

[210] Rose E. Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090* (2023).

[211] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023. CMB: A comprehensive medical benchmark in Chinese. *arXiv preprint arXiv:2308.08833* (2023).

[212] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Liu Jia. 2023. Emotional Intelligence of Large Language Models. (2023). arXiv:cs.AI/2307.09042

[213] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691* (2023).

[214] Yue Wang, Weishi Wang, Shafiq Joty, and Steven C. H. Hoi. 2021. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859* (2021).

[215] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. 2023. Exploring vision-language models for imbalanced learning. *arXiv preprint arXiv:2304.01457* (2023).

[216] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. *arXiv preprint arXiv:2306.05087* (2023).

[217] Zhuo Wang, Rongzhen Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, et al. 2023. Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. *arXiv preprint arXiv:2306.01499* (2023).

[218] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. (2023). arXiv:cs.CL/2304.04339

[219] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai Hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* 2022 (2022).

[220] Tianwen Wei, Jian Luan, Wei Liu, Shuang Dong, and Bin Wang. 2023. CMATH: Can Your Language Model Pass Chinese Elementary School Math Test? (2023). arXiv:cs.CL/2306.16636

[221] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382* (2023).

[222] Tzu-Tsung Wong. 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* 48, 9 (2015), 2839–2846.

[223] Patrick Y. Wu, Joshua A. Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057* (2023).

[224] Yiran Wu, Feiran Jia, Shaokun Zhang, Qingyun Wu, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, and Chi Wang. 2023. An empirical study on challenging math problem solving with GPT-4. *arXiv preprint arXiv:2306.01337* (2023).

[225] Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. *Advances in Neural Information Processing Systems* 35 (2022), 32353–32368.

[226] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477* (2023).

[227] Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2023. Ask Again, Then Fail: Large Language Models' Vacillations in Judgement. (2023). arXiv:cs.CL/2310.02174

[228] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? A comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841* (2023).

[229] Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility. (2023). arXiv:cs.CL/2307.09705

[230] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. (2023). arXiv:cs.CV/2306.09265

[231] Ruiyun Xu, Yue Feng, and Hailiang Chen. 2023. ChatGPT vs. Google: A comparative study of search performance and user experience. *arXiv preprint arXiv:2307.01135* (2023).

[232] Kai-Cheng Yang and Filippo Menczer. 2023. Large language models can rate news outlet credibility. *arXiv preprint arXiv:2304.00228* (2023).

[233] Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073* (2022).

[234] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. LAMM: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687* (2023).

[235] Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. KoLA: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296* (2023).

[236] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. MetaMath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284* (2023).

[237] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* (2023).

[238] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLMs Evaluations. (2023). arXiv:cs.CL/2306.04618

[239] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID- vs. Modality-based Recommender Models Revisited. (2023). arXiv:cs.IR/2303.13835

[240] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015* (2023).

[241] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. PMLR, 325–333.

[242] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. GLM-130B: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

[243] Beichen Zhang, Kun Zhou, Xilin Wei, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2023. Evaluating and improving tool-augmented computation-intensive math reasoning. *arXiv preprint arXiv:2306.02408* (2023).

[244] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609* (2023).

[245] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[246] Sarah J. Zhang, Samuel Florin, Ariel N. Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, et al. 2023. Exploring the MIT mathematics and EECS curriculum using large language models. *arXiv preprint arXiv:2306.08997* (2023).

[247] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675* (2019).

[248] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179* (2023).

[249] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005* (2023).

[250] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper LLM networks are fairer LLM evaluators. *arXiv preprint arXiv:2308.01862* (2023).

[251] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. (2023). arXiv:cs.CL/2309.01219

[252] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. SafetyBench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045* (2023).

[253] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. MMICL: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915* (2023).

[254] Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models. (2023). arXiv:cs.CL/2305.11262

[255] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[256] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934* (2023).

[257] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998* (2023).

[258] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. (2023). arXiv:cs.CL/2306.05685

[259] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197* (2022).

[260] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364* (2023).

[261] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).

[262] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528* (2023).

[263] Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. 2023. Efficiently measuring the cognitive ability of LLMs: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512* (2023).

[264] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of ChatGPT: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).

[265] Terry Yue Zhuo, Zhuang Li, Yujin Huang, Yuan-Fang Li, Weiqing Wang, Gholamreza Haffari, and Fatemeh Shiri. 2023. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex. *arXiv preprint arXiv:2301.12868* (2023).

[266] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

[267] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514* (2023).

Received 22 July 2023; revised 19 December 2023; accepted 28 December 2023