A Survey on Evaluation of Large Language Models

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, *Senior Member, IEEE,* Philip S. Yu, *Fellow, IEEE*, Qiang Yang, *Fellow, IEEE*, and Xing Xie, *Fellow, IEEE*

Abstract—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, educations, 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 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 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.

Index Terms—Large language models, evaluation, model assessment, benchmark

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 prepare for future eventualities (Khalfa, 1994). In particular, Artificial Intelligence (AI) researchers focus on the development of machine-based intelligence, as opposed to biologically based intellect (McCarthy, 2007). Proper measurement helps to understand intelligence. For instance, measures for general intelligence in human individuals often encompass IQ tests (Brody, 1999).

Within the scope of AI, the Turing Test (Turing, 2009), 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 regarded 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 (Gallant et al., 1990), 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) (Cortes and Vapnik, 1995) and deep learning (LeCun et al., 2015) 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) has incited substantial interest across both academic and industrial domains (Bommasani et al., 2021; Wei et al., 2022a; Zhao et al., 2023a). As demonstrated by existing work (Bubeck et al., 2023), the great performance of LLMs has raised promise that they could be AGI in this era. LLMs posses 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

- *Y. Chang, X. Wang, Y. Wu and Y. Chang are with the School of Artificial Intelligence, Jilin University, Changchun, China. The first two authors contributed equally.*
- *J. Wang, X. Yi, and X. Xie are with Microsoft Research Asia, Beijing, China.*
- *K. Zhu is with Institute of Automation, CAS. H. Chen is with Carnegie Mellon University.*
- *L. Yang, C. Wang, and Y. Zhang are with Westlake University, Hangzhou, China.*
- *Y. Wang and W. Ye are with Peking University, Beijing, China.*
- *P. Yu is with the University of Illinois at Chicago, IL, USA.*
- *Q. Yang is with Hong Kong University of Science and Technology, Kowloon, Hong Kong.*
- *Correspondence to: Yuan Wu (yuanwu@jlu.edu.cn) and Jindong Wang (jindong.wang@microsoft.com).*

Manuscript received April 19, 2005; revised August 26, 2015.

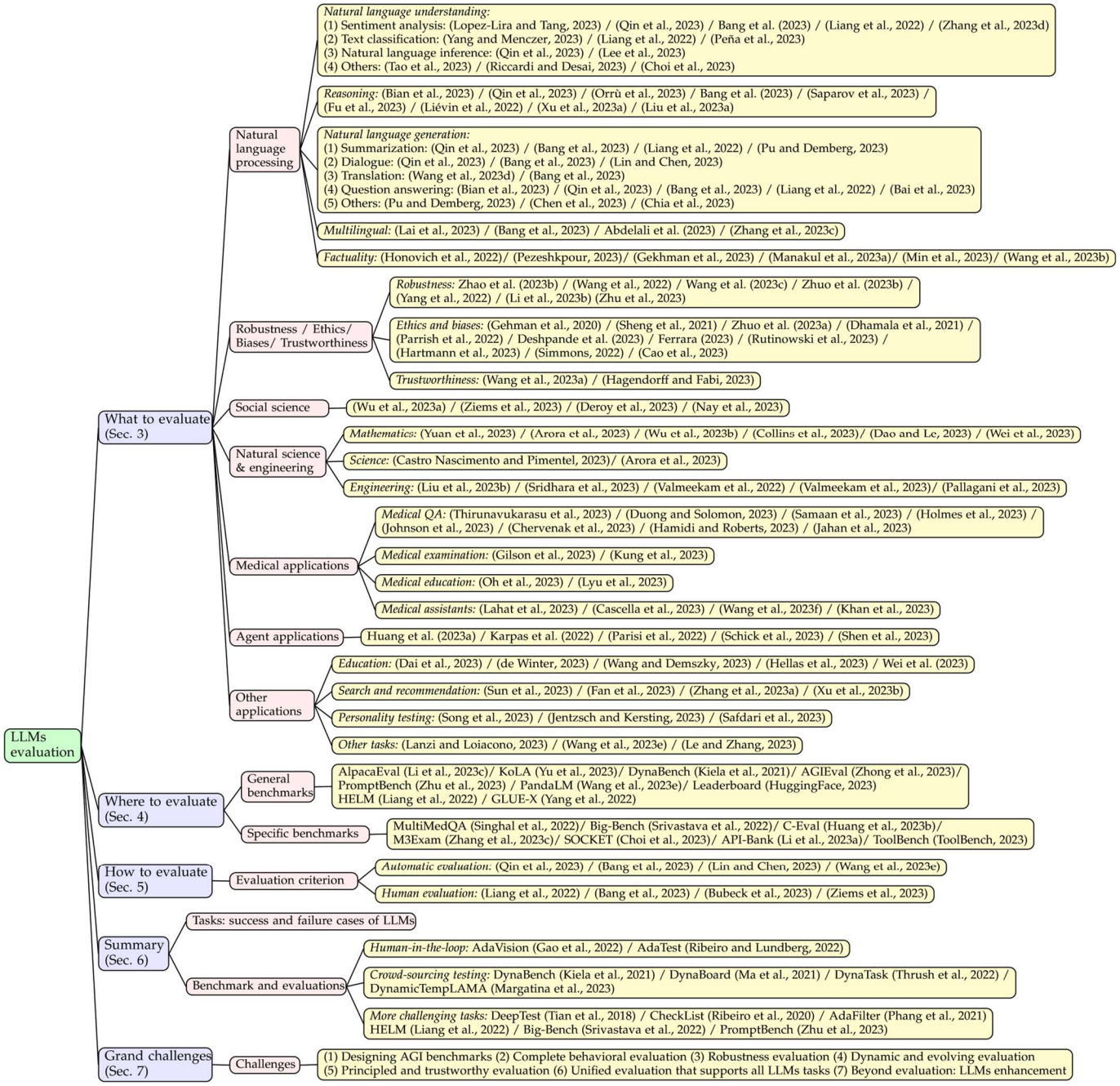

Fig. 1. Structure of this paper.

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 (Zhu et al., 2023) 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 a 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 call awareness of 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 (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b), there have been a number of research efforts aiming at evaluating ChatGPT and other LLMs from different aspects (Fig. 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 presents 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., 2023) claimed that GPT-4 can be seen as sparks of AGI, others contest this claim due to the heuristic nature of its evaluation approach.

This paper serves as the first comprehensive survey on the evaluation of large language models. As depicted in Fig. 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 (Sec. 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 Sec. 2, we provide the basic information of LLMs and AI model evaluation. Then, Sec. 3 reviews existing work from the aspects of "what to evaluate". After that, Sec. 4 is the "where to evaluate" part, which summarizes existing datasets and benchmarks. Sec. 5 discusses how to perform the evaluation. In Sec. 6, we summarize the key findings of this paper. We discuss grand future challenges in Sec. 7 and Sec. 8 concludes the paper.

2 Background

2.1 Large Language Models

Language models (LMs) (Devlin et al., 2018; Gao and Lin, 2004; Kombrink et al., 2011) 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 (Brown et al., 1992), the most

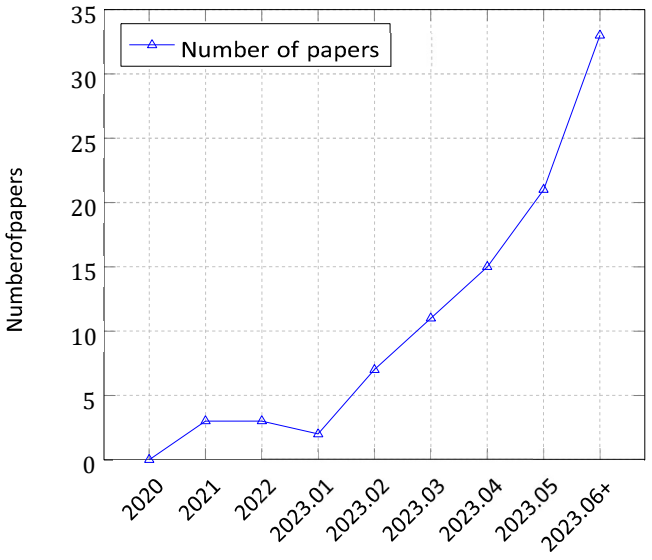

Fig. 2. Trend of LLMs evaluation papers over time (2020 - Jun. 2023, including Jul. 2023.).

common type of LM, estimate word probabilities based on the preceding context. However, LMs 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) (Chen et al., 2021; Kasneci et al., 2023; Zhao et al., 2023a) have gained significant attention in recent years due to their remarkable capabilities in natural language processing tasks. The core module behind many LLMs such as GPT-3 (Floridi and Chiriatti, 2020), InstructGPT (Ouyang et al., 2022), and GPT-4 (OpenAI, 2023b) is the self-attention module in Transformer (Vaswani et al., 2017) 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 longrange dependencies in text. One key feature of LLMs is incontext learning (Brown et al., 2020), 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) (Christiano et al., 2017; Ziegler et al., 2019) is another crucial aspect of LLMs. This technique involves fine-tuning the model using humangenerated 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 (Floridi and Chiriatti, 2020) and PaLM (Chowdhery et al., 2022), 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

TABLE 1
Comparison of traditional ML, deep learning, and LLMs

Comparison	Traditional ML	DL	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

probabilities at each position:

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

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 (Clavie et al., 2023; White et al., 2023;´ Zhou et al., 2022), 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 questionand-answer interactions (Jansson et al., 2021), 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.

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 (Berrar, 2019; Kohavi et al., 1995). 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 (Fushiki, 2011); 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 (Wong, 2015); 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.

Fig. 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 (Deng et al., 2009) and MS COCO (Lin et al., 2014) for evaluation. LLMs also use GLUE (Wang et al., 2018) or SuperGLUE (Wang et al., 2019) as the common test sets.

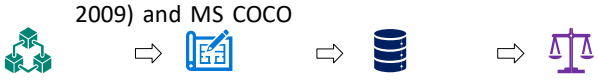

Fig. 3. The evaluation process of AI models.

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 Sec. 5.

3 What to Evaluate

What tasks should we evaluate LLMs to show their performance? On what tasks can we claim the strength and weakness of LLMs? In this section, we divide existing tasks into the following categories: natural language processing tasks, ethics and biases, medical applications, social sciences, natural science and engineering tasks, agent applications (using LLMs as agents), and others.

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.[1]

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. (2022); Zeng et al. (2022) showed that model performance is often high. ChatGPT's sentiment analysis prediction performance is superior to traditional sentiment analysis methods (Lopez-Lira and Tang, 2023) and comes close to that of GPT-
3.5 (Qin et al., 2023). In low-resource learning environments, LLMs exhibit significant advantages over small language models (Zhang et al., 2023d), but the ability of ChatGPT to understand low-resource languages is limited (Bang et al., 2023). 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.

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), Rng. (Reasoning), NLG (Natural Language Generation, including Summ. (Summarization), Dlg. (Dialogue), Tran (Translation), QA (Question Answering) and other NLG tasks), and Mul. (Multilingual tasks).

	NLU				Rng.	NLG					Mul.
Reference	SA	TC	NLI	Others		Summ.	Dlg.	Tran.	QA	Others	
(Lai et al., 2023)											✓
(Lopez-Lira and Tang, 2023)	✓										
(Bian et al., 2023)					✓				✓		
(Chen et al., 2023)										✓	
(Wang et al., 2023d)								✓			
(Yang and Menczer, 2023)		✓									
(Qin et al., 2023)	✓		✓		✓	✓	✓		✓		
(Orru et al., 2023)`					✓						
(Bang et al., 2023)	✓				✓	✓	✓	✓	✓		✓
(Liang et al., 2022)	✓ ✓					✓			✓		
(Choi et al., 2023)				✓							
(Abdelali et al., 2023)											✓
(Tao et al., 2023)				✓							
(Saparov et al., 2023)					✓						
(Lee et al., 2023)			✓								
(Lin and Chen, 2023)							✓				
(Zhang et al., 2023d)	✓										
(Fu et al., 2023)					✓						
(Pena et al., 2023)˜		✓									
(Lievin et al., 2022)´					✓						
(Riccardi and Desai, 2023)				✓							
(Bai et al., 2023)									✓		
(Zhang et al., 2023c)											✓
(Chia et al., 2023)										✓	
(Pu and Demberg, 2023)						✓				✓	
(Xu et al., 2023a)					✓						
(Liu et al., 2023a)					✓						
(Honovich et al., 2022)			✓			✓	✓			✓	
(Pezeshkpour, 2023)										✓	
(Gekhman et al., 2023)						✓					
(Manakul et al., 2023a)									✓	✓	
(Min et al., 2023)										✓	
(Wang et al., 2023b)			✓						✓		

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. Liang et al. (2022) showed that GLM-130B is the best-performed model, with an overall accuracy of 85.8% for miscellaneous text classification. Yang and Menczer (2023) 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). Pena et al. (2023) discussed the prob-˜ lem 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. (2023) 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. (2023) observed LLMs perform poorly in the scope of NLI and further fail in representing human disagreement, which indicates that LLMs still have a 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. (2023) 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 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 (2023) 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 the field of social knowledge understanding, Choi et al. (2023) evaluates how well models perform at learning and recognizing concepts of social knowledge and the results reveal 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 (Radford et al., 2018), GPT-J-6B (Wang and Komatsuzaki, 2021) and so on. This shows that supervised models significantly outperform zero-shot models and that more parameters do not guarantee more social knowledge in this setting.

3.1.2 Reasoning

From TABLE 2, it can be found that evaluating the reasoning ability of LLMs is a popular direction, and more and more articles focus on exploring its reasoning ability. The reasoning task is a very challenging task for an intelligent AI model. It requires the model not only to understand the given information, but also to reason and infer from the existing context in the absence of direct answers. At present, the evaluation of reasoning tasks can be roughly classified into mathematical reasoning, common sense reasoning, logical reasoning, professional field reasoning, etc.

ChatGPT outperform GPT-3.5 on most arithmetic reasoning tasks, demonstrating that ChatGPT has strong arithmetic reasoning ability (Qin et al., 2023), but ChatGPT still lacks mathematical reasoning (Bang et al., 2023; Frieder et al., 2023) abilities. 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 (Bang et al., 2023). In logical reasoning, Liu et al. (2023a) indicated that ChatGPT and GPT-4 outperformed traditional fine-tuning methods on most logical reasoning 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 GPT3.5 and BARD (Qin et al., 2023; Xu et al., 2023a). This is because ChatGPT is designed explicitly for chatting, so it does an excellent job of maintaining rationality. FLANT5, LLaMA, GPT-3.5, and PaLM perform well in general deductive reasoning tasks (Saparov et al., 2023). GPT-3.5 is not good at keep oriented for reasoning in the inductive setting (Xu et al., 2023a). For multi-step reasoning, Fu et al. (2023) showed PaLM and Claude2 are the only two model families that achiving 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 (Bang et al., 2023). 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 (Ott et al., 2023). In professional domain reasoning tasks, zeroshot InstructGPT and Codex are capable of complex medical reasoning tasks, but still need to be further improved (Lievin´ et al., 2022). In terms of language insight issues, (Orru et al.,` 2023) 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. 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

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 applications.

Summarization is a generation task that aims to learn a concise abstract for the given sentence. In this line of evaluation, Liang et al. (2022) showed that TNLG v2 (530B) (Smith et al., 2022) has the highest score for both scenarios, and OPT (175B) (Zhang et al., 2022) ranked second. It is disappointing that ChatGPT sometimes generates a longer summary than the input document (Bang et al., 2023). The fine-tuned Bart (Lewis et al., 2019) is still better than zero-shot ChatGPT. Specifically, ChatGPT has similar zeroshot performance to text-davinci-002 (Bang et al., 2023), but performs worse than GPT-3.5 (Qin et al., 2023). In controllable text summarization, Pu and Demberg (2023) showed that ChatGPT summaries are slightly more extractive (i.e., containing more content copied directly from the source) compared to human summaries. The above shows that LLMs, especially ChatGPT, have a general performance in summarizing tasks, but the summary and generalization ability still needs to be improved.

Evaluating the performance of LLMs on **dialogue** tasks is crucial to the development of dialogue systems and improving the 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 (Lin and Chen, 2023; Qin et al., 2023). When comparing the Claude and ChatGPT models, both models demonstrate competitive performance across different evaluation dimensions, with Claude slightly outperforming ChatGPT in specific configurations. Bang et al. (2023) test ChatGPT's for response generation in various dialogue settings: 1) Knowledge-Grounded Open-Domain Dialogue and 2) Task-Oriented Dialogue. The automatic evaluation results showed that the performance of ChatGPT is relatively low compared to GPT2 fine-tuned on the dataset for knowledge-grounded open-domain dialogue. In taskoriented dialogue, the performance of ChatGPT is acceptable, but it is prone to errors when the following problems occur: long-term multi-turn dependency, fundamental reasoning failure, and extrinsic hallucination.

While LLMs are not trained explicitly for **translation** tasks, it can indeed show strong performance. Wang et al. (2023d) showed that ChatGPT and GPT-4 demonstrated superior performance compared to commercial machine translation (MT) systems in terms of human evaluation and outperformed most document-level NMT methods in terms of sacreBLEU. When comparing ChatGPT to traditional translation models during contrastive testing, it exhibits lower accuracy. On the other hand, GPT-4 showcases a robust capability in explaining discourse knowledge, despite the possibility of selecting incorrect translation candidates. The results in (Bang et al., 2023) suggested that ChatGPT could perform X → Eng translation well, but it still lacked the ability to perform Eng → X translation. In summary, while LLMs perform satisfactorily in translation tasks, there is still room for improvement. Specifically, enhancing the translation ability from English to non-English languages should be prioritized.

Question answering is one of the key technologies in the field of human-computer interaction, and it has been widely used in application scenarios such as search engines, intelligent customer service, and intelligent question answering. Measuring the accuracy and efficiency of QA models will have important implications for these applications. Liang et al. (2022) showed that InstructGPT davinci v2 (175B) performed best in terms of accuracy, robustness, and fairness for the 9 question answering scenarios, among all the evaluated models. GPT-3.5 and ChatGPT achieve significant improvements over GPT-3 on the task of answering general knowledge questions. ChatGPT outperforms GPT3.5 by over 2% in most domains (Bian et al., 2023; Qin et al., 2023). However, ChatGPT falls slightly behind GPT3.5 on CommonsenseQA and Social IQA. This is because ChatGPT is likely to be cautious, refusing to give an answer when there is not enough information. Fine-tuned models, including Vicuna and ChatGPT, demonstrate near-perfect performance in terms of their scores, far outperforming models without supervised fine-tuning (Bai et al., 2023; Bang et al., 2023). Overall, LLMs performed flawlessly on QA tasks, and can further improve performance on social, event, and temporal commonsense knowledge in the future.

There are also other generation tasks. In the field of **sentence style transfer**, Pu and Demberg (2023) showed that ChatGPT outperformed the previous supervised SOTA model by training on the same subset for few-shot learning, as evident from the higher BLEU score. In terms of controlling the formality of sentence style, ChatGPT's performance still exhibits significant differences compared to human behavior. In **writing tasks**, Chia et al. (2023) found that LLMs perform consistently across writing-based tasks including informative, professional, argumentative, and creative writing categories, showing their general writing ability. In **text generation** quality, Chen et al. (2023) showed that ChatGPT was able to effectively evaluate text quality from various perspectives in the absence of reference texts and outperformed most existing automated metrics. Using ChatGPT to generate numerical scores for text quality was considered the most reliable and effective method among various testing methods.

3.1.4 Multilingual tasks

Many LLMs are trained on mixed-language training data. While English is the predominant language, 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, given the relatively recent emergence of this technology, LLMs are primarily evaluated on English data, while evaluating their multilingual performance is an important aspect that cannot be ignored. Several articles have provided comprehensive, open, and independent evaluations of LLMs performance on various NLP tasks in different non-English languages, offering appropriate perspectives for future research and applications.

Abdelali et al. (2023) evaluated the performance of ChatGPT in standard Arabic NLP tasks and found that ChatGPT had lower performance compared to SOTA in the zero-shot setting for most tasks. Bang et al. (2023); Lai et al. (2023); Zhang et al. (2023c) used