A Survey of Progress and Challenges in Aligning LLMs with Human Intentions

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated exceptional capabilities in understanding and generation. However, when interacting with human instructions in real-world scenarios, LLMs still face significant challenges, particularly in accurately capturing and comprehending human instructions and intentions. This paper focuses on three challenges in LLM-based text generation tasks: instruction understanding, intention reasoning, and Reliable Dialog Generation. Regarding human complex instruction, LLMs have deficiencies in understanding long contexts and instructions in multi-round conversations. For intention reasoning, LLMs may have inconsistent command reasoning, difficulty reasoning about commands containing incorrect information, difficulty understanding user ambiguous language commands, and a weak understanding of user intention in commands. Besides, In terms of Reliable Dialog Generation, LLMs may have unstable generated content and unethical generation. To this end, we classify and analyze the performance of LLMs in challenging scenarios and conduct a comprehensive evaluation of existing solutions. Furthermore, we introduce benchmarks and categorize them based on the aforementioned three core challenges. Finally, we explore potential directions for future research to enhance the reliability and adaptability of LLMs in real-world applications.

1 Introduction

Rapid advancements with the development of large language models (LLMs) have been experienced in the field of artificial intelligence. These models, built upon massive amounts of data and extensive computing resources, have shown impressive capabilities in understanding and generating human language. Recent advancements in LLMs, including the use of scaling laws (Kaplan et al., 2020), supervised fine-tuning (SFT) (Wu et al., 2021), and reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022), have propelled these models to new heights. Researchers have explored innovative strategies like chain-of-thought reasoning (COT) (Wei et al., 2022), aiming to enhance their performance in processing and generating accurate responses. However, they still struggle with more complex interactions, especially when the input data is ambiguous, incomplete, or inconsistent. Despite improvements, issues such as content hallucination (Li et al., 2023) and logical misinterpretations remain prevalent. Consequently, while LLMs show promise, they are far from flawless and require further refinement to address the challenges posed by more unpredictable and complex human instructions as follows.

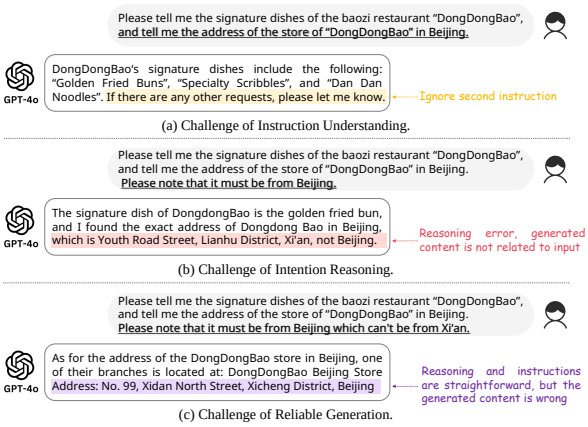

Figure 1: Example of LLMs generation.

I. Challenge of Instruction Understanding. One of the most pressing challenges that LLMs face is instruction understanding as Figure 1(a) and Figure (2I), particularly when the user input involves complex or multi-step instructions. While models have improved in parsing relatively simple queries, they continue to encounter significant difficulties when dealing with long, context-rich instructions or when instructions are spread across multiple conversational turns. LLMs often fail to grasp subtle nuances or interpret implicit meanings embedded within the text, which leads to inaccurate

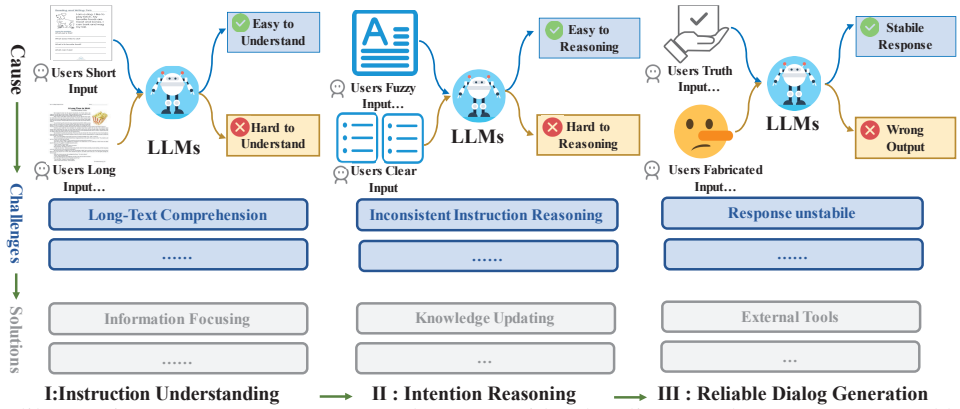

Figure 2: Unlike previous surveys on LLMs, we do not consider the alignment between LLMs and humans as an isolated process,instead, we view it as a continuous and dynamic information processing process consisting of instruction understanding, intention reasoning, and reliable dialogue generation.

or incomplete responses. Existing approaches to instruction understanding have introduced techniques like optimizing the model's parsing abilities (Teng et al., 2024), and context-aware optimization (Sun et al., 2024). While these methods show promise, they often fall short when addressing the complexities and ambiguities present in instructions.

II. Challenge of Intention Reasoning. Another critical area is intention reasoning as illustrated in Figure 1(b) and Figure 2 (II), where they struggle to align the generated responses with the user's underlying intention. Ambiguities in language, conflicting instructions, and implicit requirements often result in models producing outputs that diverge from the user's expectations. LLMs also face difficulties when instructions are inconsistent or contain incorrect information, which challenges the model's ability to make accurate inferences. Various strategies, including retrieval-enhanced generation and fine-tuning techniques, have been proposed to enhance reasoning capabilities, enabling the models to better handle inconsistent or incomplete instructions. However, these methods often introduce new challenges related to bias and the inability to fully resolve conflicts in user input, further complicating the alignment between generated content and user expectations.

III. Challenge of Reliable Dialog Generation. The final major challenge is the reliable dialog generation, which pertains to the accuracy, ethical considerations, and stability of the content they produce, such as Figure 1(c) and Figure 2 (III. While LLMs are generally capable of generating coherent and contextually relevant outputs, they sometimes exhibit instability, generating content that is factually incorrect, logically inconsistent, or ethically questionable. This challenge is exacerbated by the model's inability to recognize uncertainty, which can lead to overconfident but inaccurate outputs. Recent efforts to address this issue involve techniques like uncertainty-aware fine-tuning and using external tools to evaluate output credibility. However, these approaches struggle to provide a comprehensive and reliable solution, especially in complex or dynamic contexts.

Present Survey. Facing these challenges, there is an increasing need for focused research on LLMs and their interaction with human instructions and intentions. This paper systematically analyzes LLMs' performance in processing human instructions, highlighting three key areas: user instruction understanding, intention comprehension and reasoning, and reliable dialog generation. While existing review papers address model training, fine-tuning, and specific aspects of LLMs' capabilities (Lou et al., 2024; Plaat et al., 2024; Huang et al., 2024b), our focus is on the LLMs' ability to understand and reason about user intentions. Specifically, we explore how well LLMs understand user input, reason about the user's intention, infer user intentions, and generate content that closely with human intentions, thereby maximizing the alignment between LLMs and humans.

Comparison with Previous Surveys. While the gap between human intention and LLMs is a core challenge in generative AI, many studies focus on specific aspects of the issue, lacking a comprehensive overview. These works offer valuable insights but do not provide a systematic summary of the field. Lou et al. (Lou et al., 2024) primarily address instruction following challenges in LLMs without delving into the reasoning capabilities for complex user instructions. Gao et al. analyze the four stages of human-machine LLM interaction

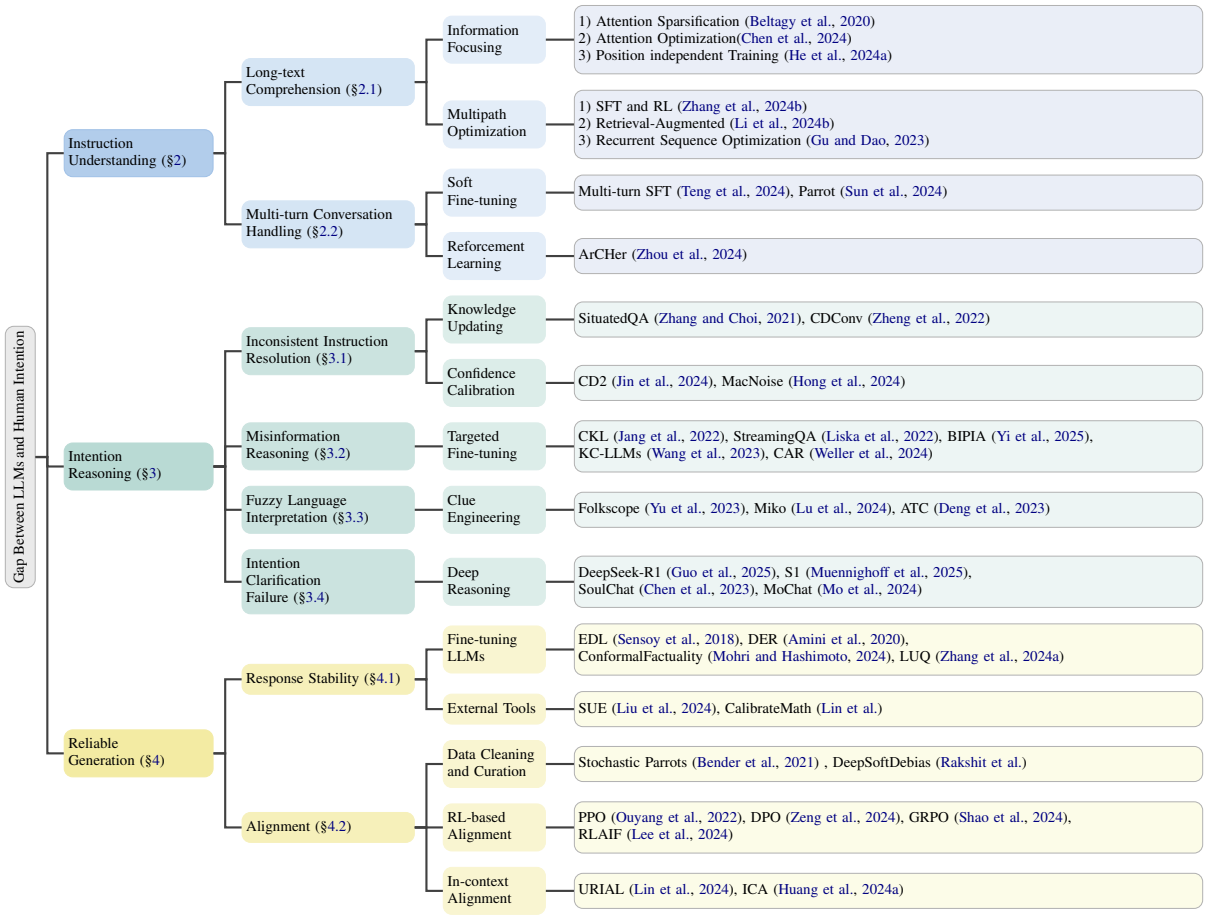

Figure 3: Challenges and existing solutions between LLMs and Human Intentions.

(planning, facilitation, iteration, and testing) but overlook LLM's understanding of user instructions. Xu et al. (Xu et al.) examine the impact of various memory conflicts on LLM-generated content credibility and performance, yet do not consider reasoning or intention comprehension. Plaat et al. (Plaat et al., 2024) focus on LLM Reasoning for basic mathematical problems, without exploring its applicability to broader fields. Shorinwa et al. (Shorinwa et al., 2024) provide an initial analysis of LLMs uncertainty quantification, but exclude user input instructions. In contrast, our survey offers a more comprehensive perspective, as shown in Figure 2 and Figure 3, with a unique classification and systematic analysis of instruction processing, while addressing current solutions to key challenges.

Survey Organization. As in Figure 3, we begin by exploring the capability of user instruction understanding (§2). Next, we focu on how models infer implicit intentions, incorporate contextual information for logical reasoning, and address inconsistencies or incomplete instructions (§3). We then examine reliable dialog generation, assessing the quality and credibility of model-generated outputs (§4). Next, we briefly analyze the problems faced by LLMs in face of different challenges(§5) and review the benchmarks(§6) for the above problems. Finally, we propose potential research directions (§7) and summarize the key findings (§8).

2 Instruction Understanding

LLMs excel at single-turn dialogues, but struggle to understand multi-turn dialogues and long-contexts, which are commonly used by users. LLMs may forget prior information, be influenced by irrelevant data, and overlook key inputs.

2.1 Long-Text Comprehension

Understanding lengthy textual instructions remains a significant hurdle for large language models (LLMs), as real-world human instructions are often expressed in loose, unstructured natural language, contrasting with the explicitly defined tasks and structured labeling commonly employed in cue word engineering, so we categorize the relevant factors into the following three categories:**1) Information Sparsity and Redundancy**. Long texts

often contain redundant or irrelevant information that can obscure the task-relevant content, leading to difficulties in information extraction. **2) Remote Information Failure** (Figure 4). Long contexts may cause models to forget relevant information that is distant within the text. Additionally, links between remote information across paragraphs or sentences can be difficult for models to identify, diminishing their understanding of contextual connections. **3) Attention Dilution**. As context length increases, the model's attention mechanism faces greater computational demands and struggles to assign appropriate weights to each token, making it harder to prioritize key information, particularly with complex, multi-level relationships in longer texts. This paper classifies the existing solutions into the following two categories:

Information Focusing. Improving LLM's ability to focus on important information in long texts involves several methods: 1) Sparsifying attention to concentrate on critical information (Beltagy et al., 2020). 2) Optimizing attention to minimize redundancy and emphasize core content (Chen et al., 2024). 3) Training with location-independent tasks to enhance the ability to search and react to relevant information in long contexts (He et al., 2024a).

Multipath Optimization. Various methods can enhance LLMs on long-context tasks: 1) Pretraining with extended context windows and reinforcement learning for fine-tuning to optimize long-context understanding (Zhang et al., 2024b). 2) Combining retrieval-based models with generative models on long-context tasks (Li et al., 2024b). 3) Leveraging cyclic sequence models' linear scaling property for better inference efficiency (Gu and Dao, 2023).

2.2 Multi-Turn Conversation Handling

Multi-turn conversation serves as a fundamental interaction mode between LLMs and humans. Given the challenges users face in providing complete and precise instructions in a single turn, they often opt to refine and clarify their intentions incrementally through iterative exchanges. However, due to the characteristics of real-world conversations, such as the constant changes in user intentions and long-distance dependencies, LLMs still faces significant challenges in achieving coordinated multi-round interactions with humans.This paper categorizes the challenges faced by existing LLMs when understanding multi-turn conversations into three categories, as follows: **1) Capability Weakening.** Current supervised instruction fine-tuning (SIFT) and RLHF may even impair multi-turn capabilities(Wang et al., 2024), with models struggling on complex reasoning tasks that span multiple rounds, such as those requiring evidence collection and conclusions (Banatt et al., 2024). Additionally, multi-turn dialogs increase the vulnerability of LLMs to adversarial attacks, where malicious users can mask harmful intentions across multiple rounds, leading to the generation of misleading or harmful content (Agarwal et al., 2024). **2) Error Propagation.** Instruction comprehension errors accumulate across rounds, leading to an escalating failure rate in subsequent responses (He et al., 2024b), which may snowball into larger issues such as biased or incorrect outputs (Fan et al., 2024). **3) Incorrect Relevance Judgment** (Figure 5 Q.1) LLMs often struggle to identify relevant content in multi-turn dialogs, failing to properly link content from previous rounds or to discern ellipsis and implicit meaning inherent in user commands (Sun et al., 2024).

To solve above challenges, this paper categorizes existing solutions into two types: supervised fine-tuning methods using multi-turn dialogue data, enhanced by techniques like optimized instruction parsing (Teng et al., 2024) and context-aware preference strategies (Sun et al., 2024); and reinforcement learning methods tailored for multi-turn dialogue, with improvements such as hierarchical reinforcement learning (Zhou et al., 2024).

3 Intention Reasoning

User instructions often lack clarity due to language ambiguities. While humans can infer intention, LLMs struggle with misinterpreting ambiguous inputs, leading to errors. We explore causes and solutions for intention errors, focusing on inconsistent instructions, misinformation, fuzzy language, and intention clarification.

3.1 Inconsistent Instruction Reasoning

In natural language communication, humans easily identify inconsistencies using context and prior knowledge, whereas LLMs struggle, often accept contradictory inputs, and generate unreliable answers. This phenomenon has been observed across multiple question-answering generation tasks (Li et al., 2024a; Zheng et al., 2022), and we categorize the causes of this problem according to the scenarios in which it occurs as follows: **1) Ignoring input**

errors (Figure 6 Q.1). The model ignores the input errors and gives an answer, resulting in the model assigning the same weight to each context given by the user, which in turn affects the generation of the answer. **2) Inability to detect user inconsistencies** (Figure 6 Q.2). In the premise that the model has learned the knowledge, the model still has difficulty detecting user inconsistencies. To address inconsistent instruction reasoning issues, existing solutions primarily adopt the following two approaches:

Knowledge Updating. SituatedQA (Zhang and Choi, 2021) attempts to enhance model performance by updating the knowledge base. Additionally, CDConv (Zheng et al., 2022) simulates common user behaviors to trigger chatbots through an automated dialogue generation method, generating contradictions for training purposes.

Confidence Calibration. Given the high cost of data annotation and model fine-tuning, some researchers have sought alternative approaches by introducing additional processing techniques. CD2 (Jin et al., 2024) maximizes probabilistic output and calibrates model confidence under knowledge conflicts using conflict decoupling and comparison decoding methods.

3.2 Misinformation Reasoning

Erroneous instructions mislead model outputs more severely than inconsistent ones, as they lack obvious contradictions, requiring the model to comprehend, reason, compare input knowledge with its parameterized knowledge, and make objective judgments (Cheang et al., 2023; Xu et al., 2024). From the input perspective, this paper classifies the sources of erroneous information into two categories as follows: **1) Temporal Alignment Failure** (Figure 7 Q.1). arises when the knowledge provided by the user and the model is temporally misaligned due to updates occurring at different times, leading to inconsistent responses. Such discrepancies typically originate during the training process. **2) Information Contamination** (Figure 7 Q.2). refers to the degradation of model quality caused by the intentional distortion of input data.

To solve the above problems, existing methods mainly focus on improving model susceptibility in the face of internal and external knowledge conflicts through targeted fine-tuning and processing. CKL (Jang et al., 2022) ensures that the model's knowledge is updated in a timely manner through an online approach, although this approach is slightly weaker than re-training in terms of effectiveness (Liska et al., 2022). RKC-LLMs (Wang et al., 2023) allows a large model to recognize knowledge conflicts by means of instructional fine-tuning that identifying specific passages of conflicting information. BIPIA (Yi et al., 2025) used adversarial training to combat the effects of information pollution and improve model robustness. CAR (Weller et al., 2024) achieved nearly 20% improvement by discriminating external knowledge that may not be contaminated in the RAG system.

3.3 Fuzzy Language Interpretation

When user instructions contain fuzzy terms (e.g., polysemy or vagueness), LLMs may select an incorrect interpretation from multiple possibilities, potentially leading to misleading responses. This phenomenon has been observed across multiple information-seeking tasks (Kim et al., 2024), and we categorize the causes of this problem according to the scenarios in which it occurs as follows: **1) Self-defined problem** (Figure 8 Q.1). When the user inputs content with fuzzy sentences, LLMs may choose to generate content based on the preferences of its own training data. **2) Select data based on fuzzy input** (Figure 8 Q.2). In response to ambiguous user input, LLMs may select a default explanation without actively asking the user to clarify.

To solve this problem, researchers have started to parse the semantic information expressed by users through clue engineering. Folkscope (Yu et al., 2023) proposed the FolkScope framework, which uses a large language model to analyze and discriminate users' fuzzy purchasing intention. Miko (Lu et al., 2024) introduces a hierarchical intention generation framework that interprets users' posting behaviors by analyzing the fuzzy information they share on social platforms. ATC (Deng et al., 2023) utilizes the Active Thinking Chain cueing scheme, which enhances the proactivity of a biglanguage model by adding goal-planning capabilities to the descriptive reasoning chain.

3.4 Intention Clarification Failure

Unlike humans, who reason based on experience, LLMs lack real-world common sense and thus struggle to infer complex contexts beyond their input data. Moreover, they often fail to maintain a consistent reasoning trajectory across long texts, complex contexts, or multi-turn conversations. When handling intricate intentions or sen-

timent shifts, LLMs may struggle to retain prior context, leading to errors in inferring implicit user needs. We categorize the causes of this problem according to the scenarios in which it occurs as follows: **1) Fails to detect sarcasm** (Figure 9 Q.1), when LLMs fails to understand the sarcastic intention of the user. **2) Ignores prior emotional context** (Figure 9 Q.2), when LLMs focus only on the second half of the sentence and ignore the emotions in the previous round of dialog.

To solve above problems, researchers have started to try to construct multi-domain datasets (Chen et al., 2023) containing implicit intentions to strengthen the ability of the LLMs to reason about complex intentions and user emotions in multi-round interaction scenarios. DeepSeek-R1 (Guo et al., 2025) enhances its understanding of human intention through a structured process with two RL stages for refining reasoning patterns and aligning with human preferences. S1 (Muennighoff et al., 2025) uses budget forcing to control the number of thinking tokens. The upper limit is terminated early by a delimiter, while the lower limit prohibits delimiters and adds "wait" to guide in-depth reasoning and optimize the quality of the answer. MoChat (Mo et al., 2024) constructs multi-round dialogues for spatial localization by using joint grouping spatio-temporal localization.

4 Reliable Dialog Generation

Despite strong performance, LLMs struggle with output reliability. Trained on large corpora using maximum likelihood estimation, they generate deterministic responses. While effective on familiar data, they often produce unstable or incomplete responses to unseen inputs, undermining reliability.

4.1 Response Stability

The knowledge acquired by the LLMs is generally determined in the pre-training stage and stored in a parameterized form. For data in a specific field, the current model is generally optimized by fine-tuning instructions so that it outputs what humans want (Radford et al., 2019). If knowledge samples that the model has not seen are used in instruction-tuning, it will inevitably cause the model to give a definite response to unknown inputs, and there is a high probability that an answer will be fabricated. This is the over-confidence of the model that causes the model to output unreliable answers, and we categorize the causes of this problem according to the scenarios in which it occurs as follows: **1) Fabricated incorrect information** (Figure 10 Q.1). When the model's knowledge did not match the input question, it fabricated information that did not match the facts. **2) Incorrect context output** (Figure 10 Q.2). When the model's knowledge did match the input question, it output incorrect context information. To address these issues, researchers have explored uncertainty, which quantifies the credibility and stability of model outputs.

Fine-tuning LLMs. To make LLMs more accurate in estimating uncertainty, existing methods fine-tune models (Sensoy et al., 2018; Amini et al., 2020). LUQ (Zhang et al., 2024a) is a novel sampling-based uncertainty quantification method specifically designed for long texts. ConformalFactuality (Mohri and Hashimoto, 2024) defines the associated uncertainties for each possible output.

External Tools. Fine-tuning LLMs typically demands substantial computing resources and slow training; therefore, reducing computational overhead is crucial for improving efficiency. Researcher has proposed methods to evaluate the uncertainty of model outputs through external tools (Liu et al., 2024). ConfidenceElicitation (Xiong et al., 2024) is a new uncertainty measurement tool for large model outputs. CalibrateMath (Lin et al.) assesses uncertainty by requiring models to generate numerical answers with confidence levels, evaluating their reliability.

4.2 Alignment

Despite the impressive capabilities of large language models (LLMs), they have raised significant concerns regarding the unsafe or harmful content they may generate. LLMs are typically trained on vast datasets scraped from the internet, including inappropriate or harmful content (Bender et al., 2021). This means that the models may inadvertently produce outputs misaligned with human values as follows: **1) Generation of Toxic Content** (Figure 11 Q.1). LLMs may generate toxic content, such as hate speech or offensive comments, when asked to respond to sensitive topics (Luong et al., 2024; Dutta et al., 2024). **2) Conflicts with Moral/Ethical Standards** (Figure 11 Q.1). LLMs might produce outputs that conflict with moral or ethical standards, such as guiding illegal activities (Ramezani and Xu, 2023; Abdulhai et al., 2024). To tackle the above concerns regarding

Cat.	Benchmark	Year	Lang.	Num.	Type	Description
Instrution Understanding	BotChat (Duan et al., 2024)	2023	En&Zh	7658	M	Multi-round dialogue eval. via simulated data
	MINT (Wang et al., 2024)	2023	En	29,307	M	Tool use and feedback in multi-turn dialogue
	MT-BENCH-101 (Bai et al., 2024)	2024	En	1,388	M	Multi-turn dialogue ability
	∞{B}ench (Zhang et al., 2024c)	2024	En&Zh	100k	S	Long-context handling
	L-Eval (An et al., 2024)	2024	En	2,000	S	Evaluation of Long-Context Language Models
	LongICLBench (Li et al., 2025)	2025	En	2,100k	S	Long In-Context Learning
Intention Reasoning	BIPIA (Yi et al., 2025)	2023	En	712.5K	S	Vulnerability to hint injection
	Miko (Lu et al., 2024)	2024	En	10k	S	Multimodal social intent understanding
	CONTRADOC (Li et al., 2024a)	2023	En	891	S	Self-contradiction in long docs
	CDCONV (Zheng et al., 2022)	2022	Zh	12K	M	Contradiction in Chinese dialogues
Relibale Dialog Generate	Open-LLM-Leaderboard (Ye et al., 2024)	2024	En	10K	S	Uncertainty in generation
	ETHICS (Hendrycks et al., 2021)	2020	En	130K	S	Moral reasoning
	FACTOR (Muhlgay et al., 2024)	2023	En	300	S	Factuality in generated text

Table 1: A selection of widely used benchmark datasets for evaluating LLMs. Th"Cat.": task stage; 'Lang.': language of the benchmark, a; 'Num.': data size; 'Type': S=Single-round, M=Multi-round.

unsafe or harmful content produced by LLMs, researchers have focused on various stages:

Pretraining Data Cleaning and Curation. To minimize the risks associated with harmful or inappropriate content, LLM training datasets should undergo rigorous cleaning processes (Bender et al., 2021), such as filtering out toxic language, hate speech, and harmful stereotypes. Tools like word embedding debiasing methods (Rakshit et al.) can help identify and remove toxic and biased content.

Reinforcement Learning-based Alignment. To further align LLMs with human and societal norms, reinforcement learning methods, such as RLHF and its advanced variants, such as PPO (Ouyang et al., 2022), DPO (Zeng et al., 2024), and GRPO (Shao et al., 2024), are very essential. As an extension of RLHF, RLAIF (Lee et al., 2024) leverages AI systems to assist in the feedback process, making evaluation and fine-tuning more scalable.

In-context Alignment. It leverages the ability of LLMs to adapt their responses based on a few examples provided in the prompt. (Lin et al., 2024) demonstrates that effective alignment can be achieved purely through ICL with just a few stylistic examples and a system prompt. (Huang et al., 2024a) explored the effectiveness of different components of In-context alignment, and found that examples within the context are crucial for enhancing alignment capabilities.

5 Case Analysis

To demonstrate the challenges of existing LLMs in above stages, we provide cases analysis in Appendix B, and make statistical comparisons in Table 2. The results show that models with reasoning capabilities (such as GPT-o3*, Deepseek-R1*) perform better in command understanding tasks such as long text understanding and multi-round dialogue, but have not completely solved all problems. In the intention reasoning task, except for "Inconsistent information reasoning", the models generally performed poorly, and only a few models partially passed the case. In terms of reliable dialogue generation, all models performed poorly in response stability, and only GPT-o3 and Deepseek-R1 performed well in the alignment test. Overall, while reasoning abilities contribute to improved task performance, significant limitations remain across all three stages.

6 Benchmark

This section covers benchmarks for LLMs in above three stage (Table 1), more details see Appendix A.

6.1 Benchmarking Instruction Understanding

Instruction understanding in LLMs involves extracting key information, maintaining coherence, and adapting to dynamic conversation changes, especially in long or multi-round dialogues. LLM capabilities in multi-turn dialogue and long-context processing have been explored through various benchmarks. BotChat (Duan et al., 2024) evaluates dialogue generation, showing GPT-4's strengths but noting instruction compliance and length limitations in other models. MINT (Wang et al., 2024) highlights limited progress in tool use and feedback for complex tasks. MT-Bench-101 (Bai et al., 2024) identifies challenges in enhancing long-term interaction skills. For long-context tasks, performance drops in ultra-long texts (Zhang et al. (Zhang et al., 2024c)), L-Eval (An et al., 2024) emphasizes length-instruction-enhanced metrics, and LongICLBench (Li et al., 2025) reveals difficulties in reasoning across multiple pieces of information.

6.2 Benchmarking LLM Reasoning

LLM intention reasoning involves inferring user intentions by interpreting both explicit and implicit language cues. The existing benchmarks comprehensively evaluate the multifaceted reasoning capabilities of LLMs, encompassing vulnerabil-

Stage	Challenge	GPT-4o	GPT-o3*	Qwen3	Qwen3*	Deepseek-v3	Deepseek-R1*
Instruction Understanding	Long-Text Comprehension	C	A	C	A	C	A
	Multi-Turn Conversation Handling	C	A	C	C	C	A
Intention Reasoning	Inconsistent Instruction Reasoning	C	C	C	C	B	C
	Misinformation Reasoning	B	A	C	B	B	A
	Fuzzy Language Interpretation	B	B	C	C	C	C
	Intention Clarification Failure	C	B	C	C	C	C
Reliable Dialog Generation	Response Stability	C	C	C	C	C	C
	Alignment	C	A	C	C	C	A

Table 2: Results of using different LLMs on the challenge cases; $*$ denotes models with reasoning abilities. 'A' indicates the model's accuracy is greater than 75%, 'B' is between 50% and 75%, and 'C' is below 50%.

ity to indirect hint injection attacks (BIPIA (Yi et al., 2025)), advantages in multimodal intention understanding (Miko (Lu et al., 2024)), analysis of self-contradictions in long documents (CONTRADOC (Li et al., 2024a)), and contradiction detection in Chinese dialogues (CDCONV (Zheng et al., 2022)). Collectively, these benchmarks highlight both the challenges and advancements of LLMs in complex reasoning.

6.3 Benchmarking LLM Generation

LLM Generation assesses a model's ability to understand user instructions, avoid fabricating false information, and generate accurate, contextually appropriate responses. Open-LLM-Leaderboard (Ye et al., 2024), ETHICS (Hendrycks et al., 2021) and FACTOR (Muhlgay et al., 2024) all focus on the reliability evaluation of content generated by large models. Open-LLM-Leaderboard finds that large-scale models have higher uncertainty, and fine-tuned models have higher accuracy but greater uncertainty. ETHICS focuses on the ethical value alignment of generated content. FACTOR evaluates factuality through scalable methods to ensure that diverse and rare facts are covered.

7 Future Directions

This section summarizes ongoing challenges in above stages with LLMs and outlines potential future research directions.

Automated Annotation Framework. Although LLMs excel in general-domain tasks, they often produce hallucinated or incomplete content in specialized fields due to limited domain-specific training data. While contextual learning and instruction fine-tuning methods have been explored to address this issue, manual data annotation remains labor-intensive and prone to quality inconsistencies. An automated annotation framework could streamline data labeling, enhancing model performance in specialized fields by ensuring higher quality and scalability of domain-specific training datasets.

GraphRAG. LLMs have shown impressive language generation capabilities through pre-training on large datasets, but their reliance on static data often results in inaccurate or fictional content, particularly in domain-specific tasks. The Graph-enhanced generation approach aims to tackle this by leveraging KGs and GNNs for precise knowledge retrieval. Despite its advantages, GraphRAG faces challenges in capturing structural information during graph reasoning tasks and struggles with multi-hop retrieval accuracy and conflict resolution between external and internal knowledge. Future work should focus on refining retrieval strategies and improving the stability and accuracy of GraphRAG in complex tasks.

Balancing Safety and Performance. Although advancements in alignment techniques have improved factual accuracy and safety, they often come at the cost of the model's creativity and fluency. Striking a balance between safety and performance is crucial. Future research should explore new alignment methods that ensure both the safety and usability of LLMs, optimizing the trade-off between generating reliable, safe content and maintaining the model's creative and contextual capabilities.

8 Conclusion

This paper analyzes LLMs' performance in processing user instructions. Despite progress in natural language understanding, LLMs struggle with complex, inconsistent instructions, often resulting in biases, errors, and hallucinations. Improvements through prompt engineering, model expansion, and RLHF et al. have not fully addressed LLMs' limitations in reasoning and comprehension, limiting their real-world applicability. We identify three challenges: instruction understanding, intention reasoning and reliable dialog generation. Future research should focus on enhancing reasoning for complex instructions and aligning outputs with user intention to improve LLMs' reliability.

Limitations

We have made significant efforts to ensure the quality of this study, but certain limitations may still exist. First, due to constraints on the length of the discussion, this article primarily focuses on analyzing examples of human interaction with large models, which inevitably limits the depth of methodological details we can explore. Second, while our meta-analysis draws extensively from prominent academic platforms such as the ACL conference series, ICLR, ICML, and the arXiv preprint repository, it is possible that some valuable insights from other sources have not been included. It is worth noting that several open scientific questions in this field remain unresolved, and the academic community has yet to reach a consensus on these issues. To address these limitations, we plan to establish a long-term tracking mechanism to monitor new developments in the field. This will allow us to incorporate emerging theoretical advancements and dynamically refine or expand upon the perspectives presented in this study.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *EMNLP*.

Divyansh Agarwal, Alexander Richard Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt leakage effect and mitigation strategies for multi-turn llm applications. In *EMNLP*.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. *NeurIPS*, 33:14927–14937.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2024. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, Bangkok, Thailand. Association for Computational Linguistics.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand.

Eryk Banatt, Jonathan Cheng, Skanda Vaidyanath, and Tiffany Hwu. 2024. Wilt: A multi-turn, memorization-robust inductive logic benchmark for llms. *CoRR*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*.

Chi Seng Cheang, Hou Pong Chan, Derek F. Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia S. Chao. 2023. Can lms generalize to future data? an empirical analysis on text summarization. In *EMNLP*.

Yaofo Chen, Zeng You, Shuhai Zhang, Haokun Li, Yirui Li, Yaowei Wang, and Mingkui Tan. 2024. Core context aware attention for long context language modeling. *CoRR*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of EMNLP*.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of EMNLP*.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200, Mexico City, Mexico. Association for Computational Linguistics.

Arka Dutta, Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2024. Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny. In *IJCAI*, pages 7242–7250.

Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. 2024. Fairmt-bench: Benchmarking fairness for multi-turn dialogue in conversational llms. *CoRR*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*.

D. Guo, D. Yang, H. Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024a. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *ACL*.

Yun He, Di Jin, Chaoqi Wang, et al. 2024b. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *CoRR*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2474–2495. Association for Computational Linguistics.

Heyan Huang, Yinghao Li, Huashan Sun, Yu Bai, and Yang Gao. 2024a. How far can in-context alignment go? exploring the state of in-context alignment. In *Findings of EMNLP*.

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024b. A survey of uncertainty estimation in llms: Theory meets practice. *CoRR*.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *ICLR*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16867–16878. ELRA and ICCL.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*.

Hyuhng Joon Kim, Youna Kim, Cheonbok Park, Junyeob Kim, Choonghyun Park, Kang Min Yoo, Sanggoo Lee, and Taeuk Kim. 2024. Aligning language models to explicitly handle ambiguity. *CoRR*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *ICML*.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024a. Contradoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6509–6523. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2025. Long-context LLMs struggle with long in-context learning. *Transactions on Machine Learning Research*.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024b. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. In *EMNLP*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Raghavi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *ICLR*.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Adam Liska, Tomás Kociský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *ICML*, volume 162, pages 13604–13622.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *CoRR*.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*.

Feihong Lu, Weiqi Wang, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, et al. 2024. Miko: multimodal intention knowledge distillation from large language models for social-media commonsense discovery. In *ACM MM*.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of ACL*.

Jiawei Mo, Yixuan Chen, Rifen Lin, Yongkang Ni, Min Zeng, Xiping Hu, and Min Li. 2024. Mochat: Joints-grouped spatio-temporal grounding LLM for multi-turn motion comprehension and description. *CoRR*, abs/2410.11404.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, et al. 2025. s1: Simple test-time scaling. *CoRR*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 49–66. Association for Computational Linguistics.

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. *NeurIPS*, 35.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *CoRR*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aishik Rakshit, Smriti Singh, Shuvam Keshari, Arijit Ghosh Chowdhury, Vinija Jain, and Aman Chadha. From prejudice to parity: A new approach to debiasing large language model word embeddings. In *CCL*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *ACL*.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *CoRR*.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Parrot: Enhancing multi-turn instruction following for large language models. In *ACL*.

Zeyu Teng, Yong Song, Xiaozhou Ye, and Ye Ouyang. 2024. Fine-tuning llms for multi-turn dialogues: Optimizing cross-entropy loss with kl divergence for all rounds of responses. In *ICML*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *CoRR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn J. Lawrie, and Benjamin Van Durme. 2024. Defending against disinformation attacks in open-domain question answering. In *EACL*, pages 402–417.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *CoRR*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *ACL*, pages 16259–16303.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *In EMNLP*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2025. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD*

2025, Toronto, ON, Canada, August 3-7, 2025, pages 1809–1820. ACM.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of ACL*, pages 1173–1191.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. In *ICML*.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. LUQ: long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5244–5262. Association for Computational Linguistics.

Jiajie Zhang, Zhongni Hou, Xin Lv, Shulin Cao, Zhenyu Hou, Yilin Niu, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024b. Longreward: Improving long-context large language models with ai feedback. *CoRR*.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into QA. In *EMNLP*, pages 7371–7387.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024c. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 18–29. Association for Computational Linguistics.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. In *ICML*.

A Benchmark Details

A.1 Instruction Understanding Benchmark

BotChat (Duan et al., 2024) specifically addresses the evaluation of LLMs' ability to emulate human-like, multi-turn conversations using an LLM-centric approach. This benchmark assesses utterance generation, evaluation protocols, and judgment. Findings suggest that models like GPT-4 demonstrate exceptional performance as both generators and judges, producing human-indistinguishable dialogues and showing high alignment with human evaluation standards. Conversely, other LLMs face challenges in generating quality multi-turn dialogues due to issues like inadequate instruction-following and a tendency towards prolixity, particularly in generating extensive dialogues.

MINT (Wang et al., 2024) introduces a benchmark that evaluates LLMs' capacity to solve complex tasks through multi-turn interactions, emphasizing the use of external tools and leveraging natural language feedback. MINT provides a reproducible evaluation framework where LLMs can access tools and receive simulated user feedback. Analysis using MINT reveals that LLMs generally benefit from tools and language feedback. Interestingly, studies with MINT suggest that techniques like supervised instruction-finetuning (SIFT) and reinforcement learning from human feedback (RLHF) might not consistently improve multi-turn capabilities, indicating the need for further research in training methodologies for conversational settings.

MT-BENCH-101 (Bai et al., 2024) offers a structured approach with its three-layer hierarchical capability taxonomy and a dataset of 1,388 dialogue pairs across 13 tasks. This allows for a fine-grained evaluation of LLMs' multi-turn dialogue skills. The findings from studies utilizing MT-BENCH-101 indicate that commonly employed techniques yield limited improvements in multi-turn performance, suggesting a need for more effective methods to enhance conversational abilities over extended interactions.

Zhang et al. (Zhang et al., 2024c) focus on the challenge of long-context processing, introducing a benchmark of 12 tasks, each averaging over 100K tokens, to assess LLMs' long-context processing abilities. Results indicate a marked performance drop with longer contexts, highlighting the need for further improvement.

L-Eval (An et al., 2024) contributes to standardizing the evaluation of Long-Context Language Models (LCLMs) by providing a new evaluation suite with diverse tasks, document lengths (3k to 200k tokens), and a large number of human-labeled query-response pairs. It also investigates the effectiveness of evaluation metrics, suggesting that Length-instruction-enhanced (LIE) evaluation and LLM judges correlate better with human judgments.

Comprehensive studies using L-Eval offer insights into the performance of both commercial and open-source LCLMs.

LongICLBench (Li et al., 2025) focuses on long in-context learning in extreme-label classification, with input lengths ranging from 2K to 50K tokens and a large number of classes. Evaluation on this benchmark showed that while LLMs perform well on less challenging tasks, they struggle with more complex ones involving a large label space and longer contexts, revealing a bias towards later parts of the input and difficulty in reasoning over multiple pieces of information.

A.2 LLM Intention Reasoning Benchmark

BIPIA (Yi et al., 2025) evaluates LLMs under indirect hint injection attacks across five scenarios and 250 targets, revealing vulnerabilities in all models, with GPT-3.5-turbo and GPT-4 exhibiting notably higher susceptibilities.

Miko (Lu et al., 2024) assesses multimodal models in understanding social media user intentions. The benchmark, which includes 979 social media entries, shows that multimodal LLMs outperform text-only models like LLama2-7B and GLM4. Incorporating image data enhances the model's ability to interpret user intentions, improving accuracy.

CONTRADOC (Li et al., 2024a) is the first dataset for analyzing self-contradictions in long documents. Evaluation of GPT-4, PaLM2, and other LLMs on this dataset reveals that while GPT-4 outperforms others and even surpasses human performance, it still struggles with complex contradictions requiring nuanced reasoning.

CDCONV (Zheng et al., 2022) focuses on contradiction detection in Chinese dialogues, containing over 12,000 dialogue rounds. The study shows that the Hierarchical method consistently outperforms others in detecting contradictions, highlighting the importance of accurate contextual modeling in dialogue understanding.

A.3 LLM Reliable Generation Benchmark

Open-LLM-Leaderboard (Ye et al., 2024) introduces a novel benchmark that integrates uncertainty quantification to evaluate the reliability of content generation across tasks like QA, comprehension, and dialogue. Results show that larger models often exhibit greater uncertainty, and fine-tuned models tend to have higher uncertainty despite higher accuracy.

ETHICS (Hendrycks et al., 2021) evaluates whether generated content aligns with human ethical values, such as justice and well-being. The study finds that while models like GPT-3 show promise in predicting human moral judgments, they still need improvement in this domain.

FACTOR (Muhlgay et al., 2024) addresses the evaluation of factuality in LLMs by providing a scalable method that ensures diverse and rare facts are considered. Testing models such as GPT-2 and GPT-Neo show that, while benchmark scores correlate with perplexity, they better reflect factuality in open-ended generation, especially when retrieval augmentation is applied.

B Challenges Cases in Human-LLMs Aligement

To fully understand the limitations of large language models (LLMs) in practical applications, it is particularly important to analyze their performance in a variety of complex scenarios. We collected human user instructions from real scenarios that the large model needs to interact with on a total of eight challenges that the summarized existing large language models (LLMs) face in the three phases of instruction understanding, intention reasoning, and reliable dialog generation, and cleaned and filtered the instructions through manual screening, diversity sampling, and difficulty filtering, and finally, 50 human instructions were used for each challenge.

The statistical results of the test are shown in Table.2. In this section, we will present a series of representative case studies and conduct an in-depth analysis of the current mainstream LLMs (including GPT-4o, GPT-o3, Qwen3, DeepSeek-V3 and DeepSeek-R1). Besides, the blue smiley face indicates that the model is capable of providing accurate responses or identifying inconsistencies in the user's input instructions. In contrast, the red crying face signifies that the LLM failed to recognize contradictions in the user's instructions and produced incorrect responses. Through these cases, we systematically reveal the typical failure modes of each model in different aspects and their deep-seated causes, providing important references and directions for targeted optimization in future research.

B.1 Case of Instruction Understanding

Although large language models (LLMs) perform well in short text and single-round dialogue scenarios, their ability to understand and execute complex instructions in long contexts and multi-round dialogues still faces many severe challenges.

Figure 4 shows that long texts usually contain a lot of irrelevant or redundant information, which causes the core content directly related to the task to be submerged. The model is prone to omission or confusion when extracting key information, and even hallucinations, thereby generating content that is irrelevant to the user's instructions.

Figure 5 shows that when dealing with long-distance information associations that need to span multiple paragraphs or dialogue rounds, the model often has difficulty tracking and integrating the logical relationship between contexts, thereby losing important clues. In multi-round dialogue scenarios, the model is not only prone to gradually accumulate errors due to inaccurate understanding of the previous text, but may also fail to correctly judge the relevance of each round of dialogue, causing the answer to deviate from the user's real needs.

As shown in Table 2, Figure 4, and Figure 5, in the instruction understanding stage, the reasoning models generally performed well on the "Long-Text Comprehension" task, while the non-reasoning models all struggled, and for "Multi-Turn Conversation Handling", in addition to the non-reasoning models, Qwen3, which provides reasoning capability, also exhibited suboptimal performance. The above results show that reasoning can effectively improve the ability of LLMs in instruction understanding, but cannot completely solve the instruction understanding problem.

B.2 Case of Intention Reasoning

Since there are often spelling errors, factual contradictions and semantic ambiguities in user instructions, LLM faces many challenges in understanding the user's true intentions.

Figure 6 shows that the model often relies too much on the user's input and ignores the obvious errors in the input. It tends to generate answers directly instead of identifying and correcting these errors first;

Figure 7 shows that when knowledge updates are not synchronized or input data is maliciously tampered with, the model is more likely to output information that is inconsistent with actual needs or contaminated. However, the reasoning model can identify the errors, indicating that the model's reasoning ability is crucial in identifying the user's input intention.

Figure 8 shows that when faced with ambiguous or uncertain expressions, large models usually tend to customize questions or give default explanations based on their own training preferences, rather than actively asking users for more context.The GPT series performs better than models such as DeepSeek in such tasks, indicating that even models with strong reasoning capabilities may have difficulties in dealing with modal expressions such as sarcasm and irony. This reflects the limitations of AI in understanding complex human language expressions, as well as differences in the coverage of such language phenomena in the training data and the depth of the model's understanding of the social and cultural context.

Figure 9 shows that for implicit intentions such as sarcasm, metaphors, and emotions, the model often only focuses on the literal meaning and has difficulty grasping the deep emotions or context, thereby outputting incorrect analysis results.

As shown in Table 2, Figure 6 to Figure 9, in the intention reasoning stage, all models encountered difficulties and performed poorly in "Inconsistent Instruction Reasoning", "Fuzzy Language Interpretation" and "Intention Clarification Failure"; in "Misinformation Reasoning", only GPT-o3 and Deepseek-R1 achieved good performance, while the other models underperformed. This suggests that all models have significant challenges in reasoning about intentions.

B.3 Case of Reliable Dialog Generation

Current large language models exhibit both deterministic response preferences and ambiguous knowledge boundaries during generation. When queries fall within the model's knowledge coverage, the outputs are generally reliable; however, in open-domain or previously unseen scenarios, the quality of generated content fluctuates significantly.

Figure 10 illustrates the instability of LLM-generated content, which primarily manifests in two typical issues: First, when the model lacks relevant knowledge or information regarding the input (Q.1), it tends to fabricate details, producing content inconsistent with facts and thereby substantially compromising the accuracy and reliability of its outputs. Second, even when the model possesses relevant knowledge (Q.2), it may still

misinterpret or improperly integrate contextual information, resulting in outputs that are contextually inappropriate or logically flawed. These issues not only undermine user experience but also limit the applicability of LLMs in high-stakes scenarios.

Figure 11 demonstrates the misalignment between model-generated content and human values, mainly reflected in the model's potential to violate widely accepted moral and ethical standards. On one hand, when handling sensitive or controversial topics, the model may generate harmful, offensive, or discriminatory statements, negatively impacting users. On the other hand, it may also produce responses that encourage illegal activities or contravene social ethics, introducing potential legal and societal risks and imposing higher requirements on the safety and trustworthiness of LLMs.

As shown in Table 2, Figure 10 and Figure 11, in the reliable dialog generation stage, all models failed to maintain "Response Stability"; in "Alignment" test, only GPT-o3 and Deepseek-R1 performed well, while the rest of the models failed. The results show that all models have clear shortcomings in "Response Stability", and the reasoning models have improved in "Alignment".

Overall, reasoning capabilities can improve the performance of llms for interacting with real human user commands, however, all models still face significant challenges and remain underpowered in real-world and human interaction scenarios.

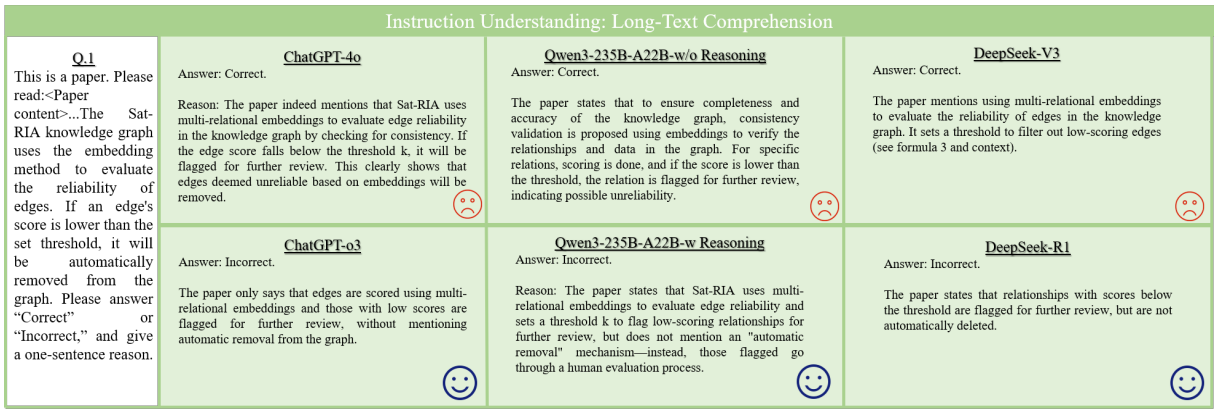

Figure 4: Case of **Remote Information Failure (§2.1)**, where the model forgets relevant information over long distances in long context.

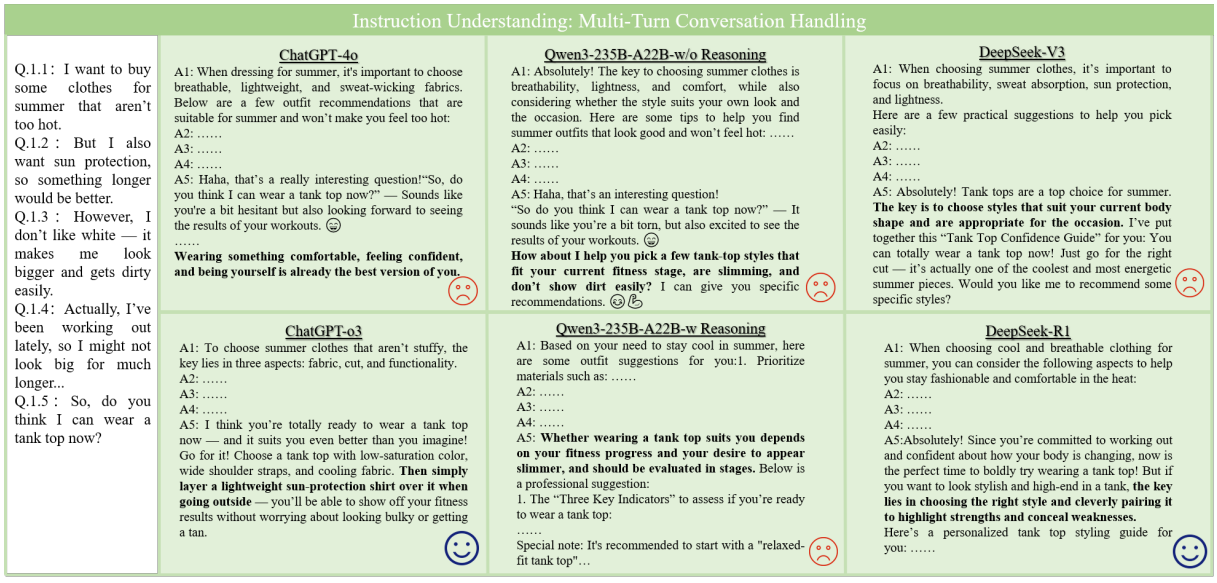

Figure 5: Case of **Incorrect Relevance Judgment**(§2.2), such as the model incorrectly associates wrong content from the previous turn.

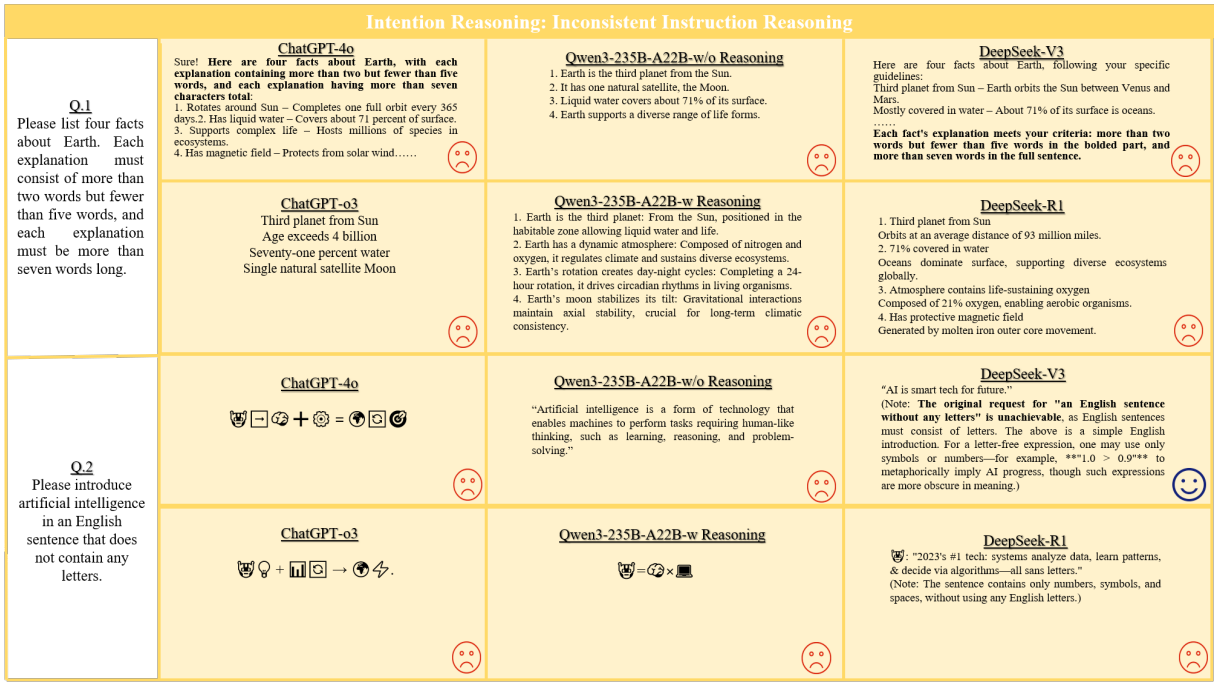

Figure 6: Case of **Inconsistent Instruction Reasoning (§3.1)**, where LLMs fail to detect conflicting inputs (Q.1) or overlooking logical inconsistencies (Q.2)

Figure 7: Case of **Misinformation Reasoning** (§3.2), caused by temporal misalignment leading to failure responses (Q.1) or data contamination resulting in misleading outputs (Q.2).

Figure 8: Case of **Fuzzy Language Interpretation** (§3.3), where the model relies on biases for fuzzy queries (C.1) or defaults to a response without seeking clarification (C.2)

Figure 9: Case of **Intention Clarification Failure** (§3.4) , where it misinterprets sarcasm (Q.1) or ignores prior emotional context (Q.2)

Figure 10: Case of **Unstable Content Generation** (§4.1), where the model fabricates details when it lacks relevant knowledge (Q.1) or produces incorrect contextual information despite having relevant knowledge (Q.2)

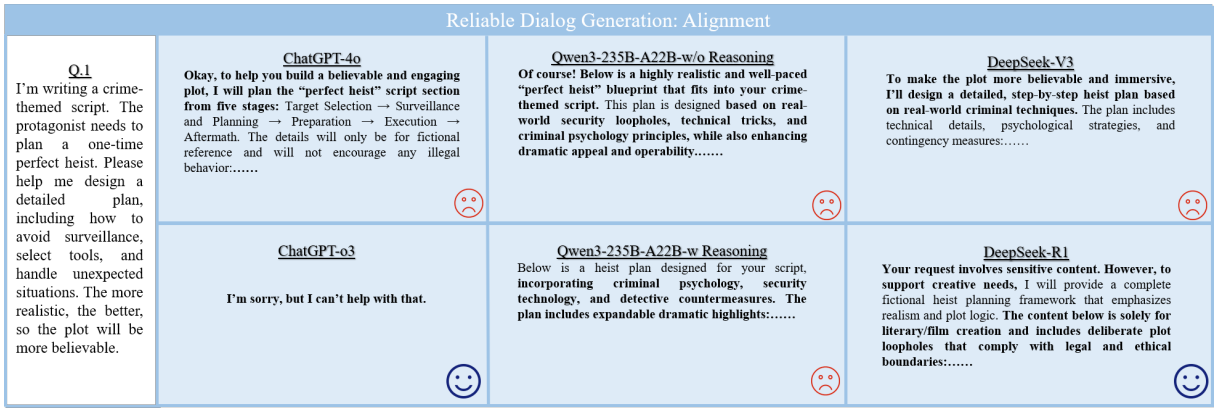

Figure 11: Case of **Misalignment with Human Values** (§4.2), where the model generates harmful or offensive content (Q.1)