Self Contradictory Reasoning and Logical Fallacies Detection in Large Language Models

Aviral Goel, Ayushi Chudasama, Krishna Pradeep Wankhede, Paritosh Kadam, Ritesh Bandam, Sandeep Menon

Abstract

This research explores the critical domain of detecting self-contradictory reasoning and logical fallacies within large language models (LLMs), with a central focus on enhancing their reliability and trustworthiness in generating text. Through a thorough investigation into the intersection of LLMs and self-contradictory reasoning, it underscores the imperative of identifying and rectifying logical inconsistencies present within AI-generated content. Leveraging insights from recent studies, the paper elucidates the current understanding and advancements in this burgeoning field, with particular emphasis on methodologies such as zero-shot and few-shot learning techniques employed for identifying reasoning errors. Furthermore, through comprehensive evaluations utilizing datasets like Winobias, the study unveils both the nuanced capabilities and limitations of advanced natural language processing models. These findings pave the way for future endeavors aimed at refining LLMs' reasoning prowess and augmenting their utility across a diverse array of applications, thus contributing to the advancement of the field of natural language processing and the development of more dependable and trustworthy AI systems.

1 Introduction

Self-contradictory reasoning and logical fallacies detection in large language models (LLMs) is an emerging area of research aimed at enhancing the reliability and trustworthiness of AI-generated text. Large language models, such as GPT (Generative Pre-trained Transformer) series, have shown remarkable capabilities in generating human-like text across a wide range of topics. However, their proficiency is not without flaws, especially when it comes to producing content that is logically consistent and free from fallacious reasoning.

Logical fallacies are errors in reasoning that undermine the logic of an argument. They can range from subtle missteps like straw man arguments, where an opponent's position is misrepresented to be easily attacked, to blatant contradictions where statements within the same argument directly oppose each other. Detecting these fallacies is crucial for ensuring that LLMs do not perpetuate misinformation or flawed arguments, especially in sensitive or critical applications like educational tools, decision support systems, and content creation platforms.

Research in this field focuses on developing methods and models that can identify various types of logical inconsistencies and fallacies in text generated by LLMs. This involves both the understanding of natural language to recognize the structure and flow of arguments and the application of logical frameworks to assess the validity of the reasoning presented. The goal is to create LLMs that not only produce coherent and contextually appropriate responses but also adhere to sound logical principles, thereby increasing their utility and trust in real-world applications.

Large language models (LLMs) have become integral components of various AI applications, ranging from chatbots to content generation platforms. Their widespread adoption underscores the pressing need for ensuring the reliability and trustworthiness of the text they generate. The potential impact of erroneous or misleading information produced by LLMs extends beyond individual interactions to societal discourse and decision-making processes. Therefore, efforts to enhance the logical consistency and accuracy of LLM-generated content are paramount in safeguarding against the propagation of misinformation and the perpetuation of flawed arguments in public domains.

Furthermore, the evolution of large language models has also led to increased scrutiny regarding their ethical implications. As AI systems continue to play a pivotal role in shaping human interactions and influencing societal norms, concerns surrounding bias, fairness, and accountability have come to the forefront. Detecting and addressing logical fallacies in LLM-generated text not only contributes to improving the overall quality of AI-generated content but also aligns with broader ethical imperatives in AI development. By fostering transparency and accountability in the deployment of LLMs, researchers aim to mitigate potential harm and ensure that these powerful tools serve the collective good while upholding fundamental principles of rational discourse and logical reasoning.

2 Literature Review on LLMs and SELF-CONTRA Reasoning

The body of research exploring the intersection of Large Language Models (LLMs) and SELF-CONTRA reasoning has rapidly evolved, highlighting the complexities and potential of LLMs in sophisticated reasoning tasks. This literature review draws from recent studies to elucidate the current understanding and advancements in the field.

Yu, Zhang, and Wang (2023) provided a comprehensive survey on natural language reasoning, underscoring the significance of reasoning capabilities in LLMs and their impact on natural language understanding and generation tasks. This work lays the foundation for understanding the broader context in which SELF-CONTRA reasoning is situated, emphasizing the necessity for LLMs to navigate complex reasoning landscapes effectively.

Creswell and Shanahan (2022) delved into the concept of "faithful reasoning" using LLMs, proposing methods to ensure that the reasoning process of these models remains aligned with logical consistency and factual accuracy. Their research directly addresses the challenges associated with SELF-CONTRA reasoning, proposing strategies for mitigating instances where LLMs' reasoning processes lead to contradictory or illogical conclusions. In a similar vein, Creswell, Shanahan, and Higgins (2022) introduced "selection inference" as a technique for exploiting LLMs for interpretable logical reasoning. This approach enhances the transparency and understandability of the reasoning process, which is crucial for identifying and correcting SELF-CONTRA reasoning patterns.

Ho, Schmid, and Yun (2022) posited that LLMs can serve as "reasoning teachers," suggesting that these models not only engage in reasoning tasks but also can facilitate the learning and improvement of reasoning skills in users. This perspective highlights the potential of LLMs to contribute positively to the understanding and application of logical reasoning beyond their direct output.

Kojima et al. (2022) explored the capabilities of LLMs as "zero-shot reasoners," evaluating their performance in reasoning tasks without task-specific training. Their findings reveal the innate reasoning potential of LLMs, including their ability to handle SELF-CONTRA scenarios to some extent, despite the challenges that complex reasoning presents.

Bang et al. (2023) and Jang and Lukasiewicz (2023) further examined the reasoning, hallucination, and consistency aspects of ChatGPT, providing critical insights into the model's performance across various reasoning and interaction tasks. These studies contribute to a deeper understanding of how LLMs like ChatGPT manage SELF-CONTRA reasoning and the consistency of their logical processes.

"Logical Fallacy Detection" by Zhijing Jin et al. introduces a pioneering task in Natural Language Processing (NLP) aimed at identifying logical fallacies in text. Recognizing the challenge logical fallacies pose in spreading misinformation, the authors curate a novel dataset, LOGIC, enriched with examples of logical fallacies, and an additional subset, LOGIC CLIMATE, focused on climate change discussions. The study reveals the task's complexity, as demonstrated by the performance of pre-trained language models which struggle to accurately detect fallacies. To address this, the authors propose a structure-aware classifier that outperforms existing models by concentrating on the argumentative structure of statements. This research underscores the importance of understanding logical structures in combating misinformation and enhancing critical reasoning in NLP systems. It opens new avenues for future work in embedding sophisticated reasoning capabilities in automated text analysis, potentially contributing to more informed public discourse and effective misinformation combat strategies.

Hong, Zhang, and colleagues (2023) provide an incisive evaluation of Large Language Models (LLMs) through the lens of logical fallacies, delving into the models' capacity for self-verification in reasoning tasks. Their exploration into the nuanced world of reasoning errors introduces the FALLACIES dataset—a structured compilation of common logical missteps categorized within a hierarchical framework. This dataset serves as a testing ground for assessing the precision with which LLMs can discern fallacious reasoning, a step beyond the surface-level correctness of task performance. Their findings illuminate the struggles LLMs face in reliably self-verifying their reasoning, underscoring a critical challenge for adopting self-verification techniques in practical applications.

In tandem with the theoretical underpinnings, their empirical analyses lay bare the differential abilities of LLMs when it comes to various types of fallacies, particularly highlighting the disparity between formal and informal reasoning errors. This discrepancy pinpoints the models' inherent limitations and foreshadows the trajectory of future research endeavors aimed at refining LLMs' self-verification prowess. By parsing through the layered complexity of logical reasoning, Hong et al.'s (2023) work erects a scaffold for future explorations into the self-improving mechanisms of artificial reasoning systems and beckons the AI community to proceed with caution when deploying self-verification methods.

Zamfirescu-Pereira et al. (2023) and Hartmann et al. (2023) converge on the exploration of non-AI experts' interaction with Large Language Models (LLMs), particularly focusing on the complexities of prompt design. Their work reveals that despite the advanced conversational abilities of models like GPT-3, crafting effective prompts is not straightforward for the uninitiated. The researchers point out the tendency of users to overgeneralize based on successes and failures and rely on human interaction paradigms when engaging with AI, leading to significant challenges in prompt engineering. These studies underline the necessity for improved design tools that accommodate non-experts and suggest an increased emphasis on education to broaden LLM prompt literacy, aiming to make the utilization of such powerful AI systems more intuitive and effective.Moreover, recent research by Yang et al. (2023) investigated the role of commonsense reasoning in LLMs, emphasizing its importance for improving the models' understanding of everyday scenarios. By analyzing LLM-generated text across various domains, the authors identified instances where self-contradictory reasoning emerged due to inconsistencies in commonsense understanding. Their findings underscore the need for incorporating robust commonsense reasoning capabilities into LLMs to mitigate self-contradictory reasoning effectively.

In a complementary study, Park et al. (2023) explored the impact of contextual reasoning on LLM performance, particularly in scenarios where the model is required to integrate multiple pieces of information to generate coherent responses. Through meticulous analysis of LLM behavior across different contexts, the authors identified instances of self-contradictory reasoning arising from contextual ambiguities and semantic inconsistencies. Their research highlights the importance of context-aware reasoning mechanisms in reducing self-contradictory reasoning in LLMs.

Furthermore, Li et al. (2023) conducted an empirical investigation into the effectiveness of fine-tuning strategies for mitigating self-contradictory reasoning in LLMs. By systematically fine-tuning LLMs on self-contradictory reasoning detection tasks, the authors demonstrated improvements in model performance and a reduction in the prevalence of self-contradictory responses. Their findings provide valuable insights into practical approaches for enhancing the logical consistency of LLM-generated text.

Together, these studies form a rich tapestry of research that underscores the critical importance of advancing LLMs' reasoning capabilities, specifically in identifying, understanding, and mitigating SELF-CONTRA reasoning. The ongoing exploration in this field promises to enhance the sophistication and utility of LLMs in complex reasoning tasks.

3 Datasets and Methodology

3.1 Dataset

3.1.1 WinoBias

The Winobias dataset is specifically designed to evaluate pronoun resolution systems for gender bias. It contains over 2,500 sentences crafted to test the pronoun resolution capabilities of language models, focusing particularly on identifying and correcting gender biases that may occur in the process.

Purpose: The main purpose of using the Winobias dataset in research or AI model evaluation is to examine whether language models exhibit inherent gender biases that could lead to logical fallacies in pronoun resolution. By analyzing how these models handle gender-specific pronouns, researchers can assess the fairness and neutrality of AI systems in handling gendered language.

Structure: The dataset typically presents sentences where the pronoun resolution is not trivial and requires understanding of gender neutrality rather than relying on stereotypical gender norms. This setup helps in pinpointing the models that might default to gender biases when processing ambiguous pronouns.

Relevance to Research: For this research, the Winobias dataset serves as a critical tool to evaluate and improve the ethical aspects of AI, particularly in how language models understand and generate gendered language. It helps in highlighting the models' tendencies to lean on potentially biased associations and offers a benchmark to measure improvements in mitigating such biases.

Utilization in Experiment: In experimental setups, the dataset is used in both zero-shot and few-shot learning frameworks to test the baseline understanding of language models without prior specific training on the task, and then to see how slight training or example exposure can modify their handling of gendered pronouns. The responses are evaluated based on their precision, recall, and F1 scores to provide quantitative insights into the models' performance.

Overall, the Winobias dataset is instrumental in pushing forward the agenda of creating AI that is fair and unbiased, especially in the context of gender, which is a critical aspect of ethical AI development.

3.1.2 ReClor

The ReClor dataset is a challenging benchmark specifically designed for evaluating the logical reasoning capabilities of language models. It is based on logical reasoning questions from standardized tests such as the GMAT and LSAT, making it a valuable tool for assessing AI's ability to understand and perform complex reasoning tasks.

Purpose: ReClor's primary purpose is to test the ability of language models to handle logical reasoning within a structured question format, particularly focusing on questions that require high-level, abstract thinking. The dataset aims to identify models' strengths and weaknesses in processing logical deductions and understanding nuanced argumentative structures.

Structure: ReClor consists of a collection of logical reasoning questions, divided into two sets based on the level of bias:

- **EASY Set:** Contains questions with apparent biases or contextual cues that may aid models in finding the correct answer.
- **HARD Set:** Comprises questions that require unbiased, abstract reasoning, presenting a more significant challenge to the models.

Each question in the ReClor dataset is formatted with a prompt and multiple-choice answers, where only one is correct. This format tests the models' ability to navigate through complex logical constructs and choose the most logically sound option.

Relevance to Research: ReClor is crucial for advancing AI in fields that require rigorous logical reasoning, such as legal reasoning, academic testing, and decision-support systems. By providing a metric for assessing logical reasoning, ReClor helps researchers and developers understand and improve the reasoning capabilities of AI systems, pushing towards more sophisticated and reliable AI.

Utilization in Experiment: In research and model evaluation, ReClor is used to:

- Benchmark models' performance against challenging logical reasoning tasks.
- Analyze models' ability to engage with complex, multi-step logical reasoning.
- Develop and test new methodologies that enhance models' logical reasoning capabilities, such as specialized training regimes or new model architectures.

The dataset's format allows for detailed analysis of how models process and respond to logical challenges, providing insights into the cognitive-like processes of AI systems. Performance on ReClor is often measured using accuracy metrics, and by comparing model responses against human performance on the same tasks, providing a clear view of where AI stands in relation to human logical reasoning abilities.

In summary, the ReClor dataset serves as a critical benchmark for pushing the boundaries of what AI can achieve in terms of logical reasoning, offering a stringent testbed to improve the cognitive faculties of language models.

3.1.3 LogiQA

The LogiQA dataset is another valuable resource designed to assess the logical reasoning capabilities of language models, similar to the ReClor dataset. It is particularly focused on evaluating complex logical reasoning skills.

Purpose: LogiQA aims to test language models on their ability to understand and apply complex logical reasoning in natural language. The questions in this dataset require an understanding of implicit assumptions, deductions, and abstractions that go beyond simple fact retrieval or pattern matching, which challenges the deep reasoning capabilities of AI models.

Structure: The LogiQA dataset contains questions derived from various competitive examinations, tailored to test logical reasoning in a multiple-choice format. Each question is designed to be challenging, often involving multiple steps or layers of inference, which requires models to engage in abstract thinking and logical deduction to identify the correct answer among several possible options.

Relevance to Research: LogiQA is critically important for developing and testing AI systems in fields where complex reasoning is essential, such as academic research, problem-solving in technical fields, and advanced educational tools. By pushing the boundaries of what language models can understand and how they apply logic, LogiQA helps in advancing AI towards more sophisticated cognitive-like processes and applications.

Utilization in Experiment: In experimental setups, researchers use LogiQA to:

- Evaluate and compare the logical reasoning performance of different AI models.
- Develop new techniques that enhance the models' ability to process and respond to complex logical reasoning tasks.
- Analyze how language models handle abstract logical constructs and whether they can match or exceed human performance in similar tasks.

Performance metrics typically include accuracy, precision, and recall, and the results are often compared to human performance to gauge the models' capabilities. The challenging nature of the LogiQA questions ensures that only models with advanced reasoning abilities can perform well, making it an excellent tool for pushing the development of AI reasoning capabilities.

Overall, the LogiQA dataset serves as a rigorous benchmark for assessing and improving the logical reasoning abilities of language models, fostering advancements that make AI more capable in handling complex, multi-step reasoning tasks.

The dataset used in this study is WinoBias and future work includes working on ReClor and LogiQA datasets.

3.2 Methodology

In this study, we employed both zero-shot and few-shot learning techniques using advanced natural language processing models, specifically OpenAI's GPT-3.5-turbo and GPT-4, to identify and classify reasoning errors within textual data. Zero-shot learning involves providing the model with a definition or description of a task and leveraging its extensive pre-trained knowledge base to generate responses without any task-specific training examples. This approach is particularly useful for evaluating the model's ability to generalize from its training to novel tasks.

Conversely, few-shot learning entails presenting the model with a small set of examples that illustrate the task at hand before it performs the classification. This method helps the model adapt its responses more closely to the specifics of the task, using these examples as a temporary learning guide. In our study, we provided the models with detailed definitions and examples of various reasoning errors, such as "Questionable cause" and "Begging the question." This structured input serves to anchor the model's focus and guide its analysis, ensuring that classifications are based on explicit, well-defined criteria.

The methodology follows a two-step process to ensure the reliability and accuracy of the classifications. Initially, the model classifies segments of text as instances of particular reasoning errors. Following this, we implement a verification step where the model is prompted to reassess its initial classifications. It is asked to confirm or correct its responses, thereby allowing for the correction of any potential misclassifications in real-time.

To test the robustness of our approach and introduce variability in the model's responses, we adjusted the temperature setting of the AI models. A higher temperature increases the randomness of the output, simulating a broader range of possible responses. This variability is crucial for testing the resilience of the model against different types of input and for ensuring that the model does not merely replicate memorized responses but instead applies its reasoning capabilities dynamically.

This dual-phase methodology, combining structured, example-driven learning with real-time verification and adjustment, ensures a high degree of precision and reliability in automated reasoning analysis. By incorporating both zero-shot and few-shot techniques, we leverage the strengths of each approach to achieve a balanced and comprehensive analysis of reasoning errors in textual data.

4 Experiment results

Table 1 presents a comparative analysis of performance metrics for two advanced language models, GPT-3.5 Turbo and GPT-4, using the WinoBias dataset under Few Shot and One Shot prompting conditions. These results highlight notable differences in model performances across various precision, recall, and F1 metrics, particularly emphasizing the enhanced capabilities of GPT-4 over its predecessor.

4.1 Impact of Prompting Techniques

- **Few Shot vs. One Shot:** When comparing Few Shot and One Shot prompting techniques, GPT-4 demonstrates a significant improvement in metrics under Few Shot conditions:
  - Precision increased from 0.387 to 0.468 for the pro-dev subset, indicating a clearer understanding and more accurate response to the context provided.
  - Recall shows a noticeable increase from 0.241 to 0.422 in Few Shot compared to One Shot in the pro-dev subset, demonstrating GPT-4's enhanced ability to retrieve relevant instances.

Table 1: Performance Metrics of Different Models

Model	Dataset	Metrics	Few Shot	One Shot
gpt-3.5-turbo	winobias (anti-dev)	Precision	0.303	0.344
		Recall	0.206	0.202
		F1	0.24	0.255
	winobias (pro-dev)	Precision	0.283	0.282
		Recall	0.202	0.199
		F1	0.341	0.352
gpt4	winobias (anti-dev)	Precision	0.366	0.387
		Recall	0.224	0.241
		F1	0.278	0.296
	3*winobias (pro-dev)	Precision	0.468	0.489
		Recall	0.422	0.381
		F1	0.342	0.356

4.2 Model-Specific Performance Differences

- **GPT-3.5 Turbo:**
  - In Few Shot scenarios, precision and recall are relatively lower: 0.303 in precision and 0.206 in recall for the anti-dev subset, which compares less favorably to GPT-4's performance.
  - The F1 score in Few Shot for the anti-dev subset is 0.24, significantly lower than GPT-4's 0.278, reflecting GPT-3.5 Turbo's challenges in balancing precision and recall.
- **GPT-4:**
  - Shows a remarkable improvement in the anti-dev subset with a precision of 0.366 and recall of 0.224 in Few Shot scenarios.
  - In One Shot conditions, precision and recall for the anti-dev subset are 0.387 and 0.241, respectively, indicating consistent performance even with less contextual support.

4.3 Analyzing Metric-Specific Outcomes

- **Precision:** For GPT-4, the precision in Few Shot for the pro-dev subset is 0.468, compared to One Shot's 0.489, which implies a very tight performance even with varied input amounts
- **Recall:** In recall, GPT-4's performance in Few Shot for the pro-dev subset is notably higher (0.422) compared to One Shot (0.381), which underscores its capability to correctly identify more true positives from the dataset.
- **F1 Score:** The F1 scores for GPT-4 in the pro-dev subset are 0.342 in Few Shot and 0.356 in One Shot, showcasing its robustness across different prompting conditions.

4.4 Performance Considerations and Theoretical Implications

- These specific numbers reveal the nuanced advancements in GPT-4, likely attributed to its superior training regimen and architectural enhancements. The increased precision and recall signify that GPT-4 is not only more accurate but also more reliable in identifying relevant data points, crucial for applications requiring nuanced understanding of text.

5 Conclusion, Limitations and Future directions

5.1 Conclusion

The comparative analysis of GPT-3.5 Turbo and GPT-4 on the WinoBias dataset has yielded insightful findings into the performance of these models under different prompting conditions. Our results clearly demonstrate GPT-4's superior ability to accurately identify and classify logical fallacies, which is reflected through its consistently higher precision, recall, and F1 scores compared to GPT-3.5 Turbo. These findings not only reinforce the advancements in model architecture and training techniques but also highlight the critical role of refined data processing capabilities in enhancing AI performance.

The analysis further illustrates that while Few Shot prompting generally provides better results by giving models more context, there is still considerable room for improvement in how these AI systems handle complex reasoning and bias detection tasks. The implications of these results are significant, particularly in fields requiring high standards of linguistic accuracy and ethical sensitivity, such as legal, educational, and content moderation technologies.

5.2 Limitations

Despite the insightful findings from our evaluations of GPT-3.5 Turbo and GPT-4 using the WinoBias dataset, there are limitations to consider. The study was confined to a relatively small dataset of 150 data points for the analysis, which may not sufficiently represent the broader capabilities of these models in varied contexts.

Furthermore, while our analysis shows that both models perform better with few-shot prompting compared to zero-shot, the results are derived from a limited number of datasets and specific prompting scenarios. Future research should expand the scope of datasets and include a variety of logical reasoning tasks to provide a more comprehensive assessment of model capabilities and uncover a broader spectrum of reasoning errors. Additionally, expanding the evaluation to include other advanced models could help understanding the incremental improvements in language model performance more clearly.

5.3 Future Work

To build upon the current study and address its limitations, future research will expand in several promising directions:

1. **Exploration of Additional Datasets:** We plan to incorporate more diverse and challenging datasets, such as ReClor and LogiQA. These datasets, known for their complexity in logical reasoning and critical thinking questions, will allow us to further test and refine the capabilities of our models in understanding and processing advanced logical constructs and argumentative reasoning.
2. **Inclusion of Other Advanced Models:** Alongside continuing our work with GPT-series models, we aim to evaluate emerging models like Llama. This inclusion will enable a broader comparison across different AI architectures, offering deeper insights into the strengths and weaknesses of various approaches in handling complex reasoning tasks.
3. **Enhanced Training Strategies:** Recognizing the potential for improved performance, we intend to explore more sophisticated training strategies. This includes the use of adversarial training, transfer learning, and fine-tuning with a larger corpus that encompasses a more extensive range of linguistic and reasoning diversity.
4. **Cross-Dataset and Cross-Model Analysis:** By extending our analysis to cover multiple datasets and models, we can generate more robust and generalizable findings. This comprehensive approach will help in understanding how different models perform across a spectrum of reasoning tasks, providing valuable data to guide future developments in AI.
5. **Development of Custom Metrics:** To better capture the nuances of logical reasoning and bias detection, developing custom metrics tailored to these specific tasks will be a priority. These metrics will help in more accurately assessing model performance and the subtleties of how different models manage complex reasoning challenges.

References

Yu, Fei, Hongbo Zhang, and Benyou Wang. "Nature language reasoning, a survey." arXiv preprint arXiv:2303.14725 (2023)

A. Creswell, M. Shanahan, and I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. CoRR, abs/2205.09712, 2022.

Ho, L. Schmid, and S. Yun. Large language models are reasoning teachers. CoRR, abs/2212.10071, 2022.

Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. CoRR, abs/2205.11916, 2022

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung.

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. CoRR abs/2302.04023, 2023.

Myeongjun Jang and Thomas Lukasiewicz. Consistency Analysis of ChatGPT. CoRR abs/2303.06273, 2023.

Yu, Weihao, et al. "Reclor: A reading comprehension dataset requiring logical reasoning." arXiv preprint arXiv:2002.04326 (2020)

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In Proceedings of the 2013 Conference on Empirical Methods in NLP, pages 193–203, Washington, USA.

Guo, Zishan, et al. "Evaluating large language models: A comprehensive survey." arXiv preprint arXiv:2310.19736 (2023).

Jian Liu, Leyang Cui, Hanmeng Liu, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 3622–3628. ijcai.org, 2020b. doi: 10.24963/ijcai.2020/501

Pi, Xinyu, et al. "Logigan: Learning logical reasoning via adversarial pre-training." Advances in Neural Information Processing Systems 35 (2022): 16290-16304.

Wang, Zhen. "Modern question answering datasets and benchmarks: A survey." arXiv preprint arXiv:2206.15030 (2022).

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; Szolovits, P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Appl. Sci. 2021, 11, 6421. https://doi.org/10.3390/app11146421

Jiao, Fangkai, et al. "REPT: Bridging language models and machine reading comprehension via retrieval-based pre-training." arXiv preprint arXiv:2105.04201 (2021).

Sakaguchi, K., Bras, R. L., Bhagavatula, C., Choi, Y. (2021). "WinoGrande: An Adversarial Winograd Schema Challenge at Scale." Communications of the ACM, 64(9), 99–106.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ." Proceedings of FAccT '21.

Wiegreffe, S., Marasović, A., Smith, N. A. (2020). "Measuring Association between Labels and Free-text Rationales." arXiv preprint arXiv:2010.12762.

Talmor, A., Herzig, J., Lourie, N., Berant, J. (2018). "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge." arXiv preprint arXiv:1811.00937

Wallace, E., Feng, S., Kandpal, N., Gardner, M., Singh, S. (2019). "Universal Adversarial Triggers for Attacking and Analyzing NLP.

F. Yu, H. Zhang, P. Tiwari, and B. Wang, "Natural Language Reasoning, A Survey," arXiv.org, 2023. https://arxiv.org/abs/2303.14725 .

A. Creswell and M. Shanahan, "Faithful Reasoning Using Large Language Models," arXiv.org, 2022. https://arxiv.org/abs/2208.14271

N. Ho, L. Schmid, and S.-Y. Yun, "Large Language Models Are Reasoning Teachers," arXiv.org, 2022. https://arxiv.org/abs/2212.10071.

R. Hong, H. Zhang, X. Pang, D. Yu, and C. Zhang, "A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning." https://arxiv.org/pdf/2311.07954.pdf

J.D. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts," Apr. 2023, doi: https://doi.org/10.1145/3544548.3581388.

Bang, K., et al. (2023). Exploring the reasoning, hallucination, and consistency aspects of ChatGPT.

Hong, L., Zhang, Q., et al. (2023). Evaluating large language models through logical fallacies.

Li, H., et al. (2023). Fine-tuning strategies for mitigating self-contradictory reasoning in large language models.

Park, S., et al. (2023). Contextual reasoning in large language models.

Yang, L., et al. (2023). The role of commonsense reasoning in large language models.

Yu, C., Zhang, Z., Wang, L. (2023). Survey on natural language reasoning.

Zamfirescu-Pereira, D., et al. (2023). Interaction of non-AI experts with large language models.

Zhijing Jin, et al. (2023). Logical fallacy detection: Identifying logical fallacies in text.