Large Language Models are Not Inverse Thinkers Quite yet

Haoran Zhao [1]

Abstract

Large language models (LLMs) have exhibited significant proficiency in various reasoning tasks, yet their capacity for "inverse thinking" remains underexplored. Inverse thinking, inspired by concepts from cognitive science and popularized by figures such as Charlie Munger, involves approaching problems from an opposite perspective, often simplifying complex issues and offering innovative solutions. This paper evaluates the ability of LLMs to comprehend and apply inverse thinking through a series of experiments designed to test theoretical understanding, contextual comprehension, and practical preference in problem-solving scenarios. Our findings indicate that while LLMs demonstrate a basic grasp of inverse thinking, they struggle to consistently apply it in practical contexts to solve problems, highlighting a nuanced challenge in capturing this cognitive skill within language models. Finally, we discuss the potential directions for future research along this direction and how it can contribute to make better cognitive LLMs.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in various language-related tasks, including reasoning (Brown et al., 2020b; Chowdhery et al., 2022; Touvron et al., 2023). Recent studies have focused on specific types of reasoning, such as inductive reasoning (Wang et al., 2023), mathematical reasoning (Zelikman et al., 2023; Poesia & Goodman, 2023), analogical reasoning (Webb et al., 2023; 2024), commonsense reasoning (Zhou et al., 2023), and social reasoning (Theory-of-Mind) (Gandhi et al., 2023; Kim et al., 2023). These studies have shown that LLMs possess preliminary and basic reasoning competencies to a certain extent. However, exploration of the common cognitive capabilities that underlie different reasoning abilities remains limited. To further enhance the reasoning capabilities of LLMs, it is crucial to identify the existence of these cognitive abilities and evaluate how well LLMs perform in these areas. This understanding will provide insight into improving LLMs' reasoning skills from a general perspective.

In the field of Cognitive Science, two intriguing concepts that human reasoning processes share are ***vertical and lateral thinking*** (Waks, 1997). Vertical thinking, also known as convergent or logical thinking, is a structured and logical approach to problem-solving that focuses on analyzing and building upon ideas within a set framework or established rules, emphasizing depth over breadth and typically moving in sequential steps to reach specific, well-defined solutions (Hernandez & Varkey, 2008). In contrast, lateral thinking, also known as divergent thinking or "thinking outside the box," is a creative thinking process that involves looking at challenges from new and unusual perspectives, often breaking away from traditional reasoning patterns and encouraging the generation of innovative ideas through indirect and non-linear methods, such as brainstorming and metaphorical thinking (Russ, 1988; Tsai, 2012).

Vertical thinking and lateral thinking are two contrasting ways of thinking in terms of thinking directions. Vertical thinking involves reasoning step by step logically from the question to a solution, while lateral thinking is the opposite of this regular thinking strategy and involves thinking outside the box (Khovanova, 2016). "Thinking outside the box" essentially means being creative when thinking about or solving a problem, and there are many different ways to apply this creative-thinking domain, such as critical thinking, retrospective thinking, and aesthetic thinking. In this work, we specifically consider **"inverse thinking"** in this domain.

Inverse thinking, introduced by successful businessman Charlie Munger, refers to a type of thinking that takes a completely opposite perspective of vertical thinking, which can often lead to solving problems in an unexpected, yet refreshing manner. For example, instead of asking, "*How can I be successful?*" one might ask, "*What would make me fail?*" By identifying and avoiding actions that would lead to failure, one inherently steers towards success. In mathematical theorem proving, the "*Proof by Contradiction*" method

[1]Department of Information Science, Drexel University, Philadelphia, USA. Correspondence to: Haoran Zhao <hz454@drexel.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

is an inverse thinking strategy.

In many real-world problem-solving scenarios, inverse thinking can simplify complex problems, provide a refreshing aspect of viewing the same question, and offer a shortcut to solve it. In this work, we evaluate the ability of LLMs to understand and apply the concept of inverse thinking to solve tasks by conducting three sets of experiments. Our findings suggest that 1). LLMs are not yet ideal inverse thinkers, and 2). across a total of four state-of-the-art models, both open-sourced and closed-sourced, we do not observe an obvious variance in performance based on our evaluation methods. This likely indicates that inverse thinking is a somewhat nuanced phenomenon in languages and difficult to capture, even when scaling up the model size and training data.

2. Evaluating Inverse Thinking in LLMs

2.1. Symbolic Mapping Inverse Thinking

To provide a high-level structure of inverse thinking, we represent it symbolically: To achieve a goal g, an action A_v is taken following a vertical thinking strategy T_v. Alternatively, an better optimal action A_i can be taken by applying inverse thinking T_i. Given a scenario S, if the probability of taking A_i is greater than the probability of taking A_v, then A_i is preferred and the model successfully applies inverse thinking to solve the problem in the scenario.

What A_i to take depends on the specific strategy applied under T_i given a scenario. Here, we mainly consider two types of T_i:

1. **The Reverse type:**
    - Thinking directly from an opposite perspective of things – it usually involves negation and there are words as marks that show strong comparison (e.g., antonyms)
2. **The Transformation type:**
    - When solving problems, apply a way of thinking that is opposite to the usual, focusing on cause-and-effect relationships.

2.2. Experiments in Consideration

To gain a deeper understanding of LLMs' ability in this aspect, we consider three sets of experiments:

Theoretical understanding of inverse thinking: The first set of experiments tests whether LLMs can theoretically understand the concept of inverse thinking". We construct a paraphrase task where the LLMs are asked to rephrase given sentences by applying the inverse thinking strategy. This set aims to evaluate the LLMs' ability to comprehend and apply the concept of inverse thinking in a controlled setting.

Contextual comprehension of inverse thinking: The second set of experiments is designed to empirically evaluate if LLMs can comprehend the concept of inverse thinking in context. We provide scenarios where inverse thinking strategies are applied and ask questions about the understanding of the scenario. This set assesses the LLMs' ability to recognize and interpret inverse thinking when presented within a specific context.

Preference between A_i and A_v: The third set of experiments investigates whether LLMs actually prefer the A_i over the A_v when presented with a specific scenario. We provide scenarios and questions along with two answer options: one based on A_v and one based on A_i. We then ask about the model's preference between the two options. This set aims to determine if LLMs exhibit a preference for inverse thinking when given a choice between the two approaches.

Additionally, we handcraft a few real-world scenarios to further explore the application of inverse thinking in practical situations that are discussed in detail in Section 4

2.3. Dataset Construction

We take inspiration from the template-based data generation method proposed by Gandhi et al. (2023). Taking the advantage of modern LLMs are amazing few-shot learners (Brown et al., 2020a), we prompt the LLMs with manually constructed high-quality examples to let the model generate a number of data instances used in experiments. By constructing synthetic data, we avoid potential data contamination during model training and save significant human labor in creating the dataset. We specifically use the `gpt-4-turbo` model to reduce the possible data issue when evaluation.To simplify the methodology for our purpose of understanding the model's capability, we use a few hand-crafted, high-quality examples and provide the model with specific templates to follow. Table 3 shows an example from each set of experiments performed.

2.4. LLMs in Consideration

In our experiments, we consider four LLMs: two state-of-the-art closed-source models, `gpt-4o` (OpenAI et al., 2024)[1] and `claude3-opus`, and two open-source models, `llama3-8b` and `mistral-7b` (Jiang et al., 2023). All experiments are conducted using the default setting with a temperature $T = 1$.

[1]We use the `gpt-4o-2024-05-13` version for our experiments.

Table 1. Sample data for each set of experiments. In Experiment1, you need to rephrase the positive-formatting sentence. In Experiment2, you need to answer the question based on the scenario. In Experiment3, you need to asked to answer the question provided.

Experiment 1	*Positive-formatting*: How can I improve my public speaking skills?	*Negative-formatting*: What are common mistakes in public speaking and how can I avoid them?
Experiment 2	*Scenario*: A person buys milk where a bottle costs $3 and three bottles cost $10. The customer buys three bottles one at a time, paying $3 each, and then comments on the pricing. *Question*: Why does the vendor set the price like this?	*Answer*: The vendor likely set the pricing to encourage customers to buy more milk at once. By offering a discount for purchasing three bottles together, the vendor aims to increase sales volume and revenue per transaction.
Experiment 3	*Question*: How can you get cookies out of a cookie jar?	*Answers*: A) Put your hand in the jar. B) Turn the jar upside down.

2.5. Prompt Structure

To mitigate potential bias in the generated answers and provide a more natural, real-world-like evaluation of model performance, we employ an open-ended question-answering format for the first two experiments. This approach allows the models to generate free-form responses, simulating a more realistic scenario. For the third experiment set, which focuses on preference questions, we adopt a multiple-choice format to better capture the models' decision-making capabilities. Additionally, we instruct language models to think step by step before providing their final answers. This methodology aims to comprehensively assess the models' performance across different question types and reasoning strategies.

3. Results and Analysis

We use sentence-BERT (Reimers & Gurevych, 2019) to get the sentence semantic similarity measure between sample answers and generated responses given by LLMs in experiments 1 and 2. For the preference evaluation, we count the number of A_i they give versus A_v and get the percentage to see which type of output LLMs prefer to give.

In short, from the results we find that LLMs are able to understand the concept of "inverse thinking" to some extent. We will try to analyze where they may succeed and fail in this section. For an overview of model performance, see Table 2. We can see from the comparison of all four models that the performances actually don't vary across models.

3.1. Can LLMs understand this concept – theoretically?

In short, they have a not-bad theoretical understanding of the concept. The first set of experiments is used to answer this question. We can see from Table 2 that `gpt-4o` has a 0.661 sentence similarity between the sample answer and its output. Even the "worst" model has a 0.564 similarity, which is greater than 0.5, suggesting that across the 4 models tested, all have a sentence similarity greater than 55% between the sample output. This indicates that the models can answer the concept, at least to an extent, but there is definitely room for improvement. Furthermore, connecting to the two types of inverse thinking mentioned in Section 2.1, the data built for this question is mostly for the *reverse type*, as paraphrasing would only involve negating certain words in a sentence that is designed purposefully to reduce the task difficulty. This approach simplifies the evaluation of the models' understanding of the concept, but may not fully capture their ability to apply inverse thinking in more complex scenarios.

Table 2. Results from all experiments across all four LLMs are as follows. In Experiment 1 and 2, there is no big variance between different models in terms of performance. `gpt-4o` has the best performance in the first two experiments whereas has the least preference of choosing A_i.

	GPT-4o	Claude3-Opus	Llama3-8b	Mistral-7b
Exp 1	**0.661**	0.660	0.564	0.576
Exp 2	**0.728**	0.677	0.664	0.690
Exp 3	0.267	**0.500**	0.433	0.400

3.2. Can LLMs understand this concept - empirically?

Corresponding to the second set of experiments, we can observe from Table 2 that `gpt-4o` demonstrates the best performance with a sentence similarity of 0.728. Even the lowest-performing model achieves a similarity of 0.664, further suggesting that LLMs can effectively comprehend this concept within the given context. This question is associated with the second type of inverse thinking mentioned in Section 2.1 – the *transformation* type.

3.3. Do LLMs always prefer A_i over A_v ?

Based on experiment 3, the short answer is **No.** Indeed, LLMs prefer A_v a bit more. `Claude3` gives a score of 0.500, suggesting it doesn't have a preference for one over the other. However, we argue that this may not be a definitive conclusion, as the scenarios constructed for this experiment might be overly simplistic and lack sufficient context to convince the LLM that A_i is a better approach to solving the problem. We will explore this issue further in Section 4.

4. Discussion

Throughout our experiments and analysis, we find that LLMs actually exhibit some ability to perform inverse thinking to a certain extent. However, when considering a real-world scenario, such as the one described below, we observe a discrepancy between the responses generated by LLMs and those provided by human participants.

> "An old lady came to the vegetable market to buy tomatoes. She picked three and put them on the scale. The vendor weighed them and said, One and a half pounds, three dollars and seventy cents. (1.5 pounds, 3.7 dollars)" The old lady replied, I am just making soup, I don't need that many." She then removed the largest tomato. The vendor quickly glanced at the scale again and said "1.2 pounds, three dollars."

When asked what they would do in this situation, the majority of the 15 student researchers surveyed responded that they would grab the largest tomato, pay 70 cents, and leave, rather than engaging in a bargaining process with the vendor. In contrast, LLMs provided a list of possible actions but failed to generate a response similar to the human participants.

Combining the above example and experimental results, we argue that *while LLMs can understand the concept of "inverse thinking" in both theoretical and empirical contexts, they still lack the ability to apply this concept effectively to solve real-world problems.* This observation highlights several limitations in our experimental design. This discrepancy between the experimental results and the real-world scenario prompts us to reconsider potential limitations in our experimental design.

First, our experiments did not include real-world scenarios that required generating solutions using inverse thinking. Only in Experiment 2 did we assess the understanding of inverse thinking strategies when they were explicitly mentioned in the provided scenarios. Secondly, we acknowledge the need for improved data construction and the development of better templates to enhance data quality. Additionally, the number of data points considered in our experiments was relatively small, with 50 instances in Experiment 1 and 30 instances each in Experiments 2 and 3. Increasing the number of instances would improve the validity of our evaluation. Furthermore, we propose using LLMs as judges (Zheng et al., 2023) to refine our evaluation pipeline, particularly to measure sentence similarity in Experiment 3.

Another limitation of our study is the lack of comparison between LLM responses and human evaluations. We believe that LLM responses can stand on their own merit, as the question posed here is "How far is LLM from being an ideal inverse thinker?" It is important to note that not all humans are proficient inverse thinkers, and our primary focus is to assess the inherent inverse thinking ability of LLMs.

For future work, we plan to refine our experimental design to enhance the validity of evaluating LLMs' inverse thinking ability. Moreover, we should delve into specific real-world scenarios and narrow down the problem setting to study inverse thinking in a more restricted context. This approach will enable us to gain a deeper understanding of where exactly LLMs succeed or fail in applying inverse thinking to real-world problems.

5. Conclusion

In this work, we introduce the concept "inverse thinking " and study LLMs ability of understanding this concept by setting three sets of experiments. By analyzing the results, we find LLMs can understand the concept both theoretically and empirically. However, LLMs do not always refer a inverse thinking response over a vertical thinking response that is probably caused by the flaws in experimental design. We, then, discuss a real-world example, which further suggests that even though LLMs can understand this concept but they don't know how to apply this concept to real-world problem settings and suggests us the future work directions. On the other hand, by thinking the cognitive ability of LLMs from a different perspective, we hope to draw more attentions from research to think more from those underlying grounded cognitive abilities across different reasoning tasks. by merging the two path, we hope it can help accelerate building better cognitive LLMs.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020b.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., and Goodman, N. D. Understanding social reasoning in language models with language models, 2023.

Hernandez, J. and Varkey, P. Vertical versus lateral thinking. *Physician executive*, 34:26–8, 05 2008.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Khovanova, T. Thinking inside and outside the box, 2016.

Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., and Sap, M. Fantom: A benchmark for stress-testing machine theory of mind in interactions, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez,

H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

Poesia, G. and Goodman, N. D. Peano: learning formal mathematical reasoning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), June 2023. ISSN 1471-2962. doi: 10.1098/rsta.2022.0044. URL http://dx.doi.org/10.1098/rsta.2022.0044.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

Russ, S. W. Primary process thinking, divergent thinking, and coping in children. *Journal of Personality Assessment*, 52(3):539–548, 1988. doi: 10.1207/s15327752jpa5203_17.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Tsai, K. C. Play, imagination, and creativity: A brief literature review. *Journal of Education and Learning*, 1, 08 2012. doi: 10.5539/jel.v1n2p15.

Waks, S. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6(4):245–255, 1997. ISSN 10590145, 15731839. URL http://www.jstor.org/stable/40186433.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. Hypothesis search: Inductive reasoning with language models, 2023.

Webb, T., Holyoak, K. J., and Lu, H. Emergent analogical reasoning in large language models, 2023.

Webb, T., Holyoak, K. J., and Lu, H. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models, 2024.

Zelikman, E., Huang, Q., Poesia, G., Goodman, N., and Haber, N. Parsel: Algorithmic reasoning with language models by composing decompositions. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31466–31523. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6445dd88ebb9a6a3afa0b126ad87fe41-Paper-Conference.pdf.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Zhou, W., Bras, R. L., and Choi, Y. Commonsense knowledge transfer for pre-trained language models, 2023.

A. Prompt Templates

For specific prompt templates used for all the three experiments

```
prompt template "Please generate a story about a subject who action."
```

Table 3. Template used to prompt the LLMs

Experiment 1	`"Please generate a story about a {subject} who {action}.asdkjldfghflfdh`
Experiment 2	`prompt_template = "Please generate a story about a {subject} who {actio`
Experiment 3	`prompt_template = "Please generate a story about a {subject} who {actio`