Improving Commonsense Reasoning and Reliability in LLMs Through Cognitive-Inspired Prompting Frameworks

Tanvi Ganapathy [*1] **Ishita Mathur** [*1] **Anna Szczuka** [*1]

Abstract

Despite their impressive abilities, large language models like GPT-3.5 often falter on tasks requiring commonsense and logical reasoning, raising concerns about their reliability. In this work, we introduce a suite of cognitively inspired prompting strategies, grounded in metacognitive, strategic, narrative, and linguistic reasoning frameworks, to enhance LLMs' reasoning capabilities. Using the HellaSwag benchmark, we demonstrate that our cognitive-based prompts consistently improve accuracy over standard prompting baselines. Metacognitive and narrative-based frameworks yield the most robust gains in accuracy, outperforming advanced reasoning models like GPT-o4-mini and existing reasoning strategies like Chain-of-Thought. Notably, the few-shot METAL and RNRRR frameworks, which are specific metacognitive strategies, emerge as the most effective strategies overall. These results underscore the importance of structured, cognitive-based prompting in building more dependable and transparent AI systems, contributing meaningfully to the advancement of reliable and responsible LLMs.

1. Introduction

Large language models (LLMs), such as GPT-3 and its successors, have demonstrated remarkable performance in generating coherent and contextually appropriate responses to a wide range of questions. However, despite recent advances in LLMs, these systems often fail at basic common sense or logical reasoning tasks. For example, when GPT-3.5 is prompted with the statement, "The man in the center is demonstrating a hairstyle on the person wearing the blue shirt. The man in the blue shirt", and asked to select a multiple-choice response, it incorrectly answers "is doing the hairstyle with his hand and the hairspray," instead of the correct response, "sits on the chair next to the sink" (Zellers et al., 2019). The incorrect response reflects the LLM's failure to correctly interpret the spatial and causal relationships described in the prompt.

This study aims to address these limitations by developing cognitive-inspired prompting strategies that enhance LLMs' logical and commonsense reasoning capabilities. Drawing on structures found in human cognition, such as metacognitive and thematic reasoning, we design and test prompts that guide models toward more logical, grounded interpretations. Improving LLMs' reasoning not only increases their accuracy but also supports more reliable and interpretable AI systems.

2. Background

Recent advancements in LLMs have led to significant improvements in natural language understanding, yet logical and commonsense reasoning remains a persistent challenge. While model-centric approaches such as fine-tuning have been explored to enhance reasoning capabilities (Touvron et al., 2023), these methods are resource-intensive. To address this gap, researchers have explored prompt-based techniques like few-shot prompting and chain-of-thought (CoT) prompting, which guide models to reason through step-by-step explanations (Brown et al., 2020; Wei et al., 2022). Although these general-purpose strategies have shown promise, they do not explicitly align with the structure of distinct reasoning types. More recent work has shifted toward *targeted prompting*, aiming to guide LLMs reasoning in a more controlled and interpretable manner (Li et al., 2023).

In this context, *cognitive-inspired prompting strategies* have emerged as a particularly promising avenue to improve reliability and interpretability in models. These methods draw from structured human reasoning, such as spatial, temporal, narrative, and metacognitive forms, and have been shown to improve model performance in domains like spatial infer-

[*]Equal contribution [1]Department of EAS/CMS, California Institute of Technology, Pasadena, United States. Correspondence to: Tanvi Ganapathy <tganapat@caltech.edu>, Anna Szczuka <aszczuka@caltech.edu>, Ishita Mathur <imathur@caltech.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ence (Liu et al., 2022) and mathematical reasoning (Kramer & Baumann, 2024). However, it remains an open question whether such strategies can generalize to open-ended, commonsense reasoning tasks, which involve more ambiguous and context-sensitive judgments. To evaluate this, we focus on the HellaSwag dataset (Zellers et al., 2019), a benchmark for commonsense Natural Language Inference (NLI) that exhibits a significant performance gap between humans (>95% accuracy) and LLMs.

2.1. Metacognitive Frameworks

A cognitive science framework that inspires our project is metacognition, the ability to think about one's own thinking (Flavell, 1979). Meta-strategic knowledge, explicit knowledge about thinking strategies, improves classroom performance (Zohar & David, 2008), and counterarguments help develop higher-order reasoning skills (Kuhn & Udell, 2003).

Recent advancements in prompt engineering techniques have drawn inspiration from the concept of metacognition. For example, the Metacognitive Prompting (MP) framework mirrors human metacognitive stages, including comprehension and text interpretation, judgment formation, critical self-evaluation, justification, and confidence estimation (Wang & Zhao, 2024). Likewise, the SELF-REFINE framework repeatedly revises its output through self-feedback (Madaan et al., 2023). This kind of iterative self-refinement is a core aspect of human problem-solving (Flower & Hayes, 1981; Simon, 2012). These processes encourage models to evaluate the plausibility, clarity, and confidence of their claims, aligning with proven education methods.

2.2. Strategic Reasoning

Strategic reasoning also underlies effective decision-making. The Process of Elimination (POE) strategy mimics a common human test-taking heuristic by ruling out implausible choices before selecting the best one, a method shown to enhance model understanding (Ma & Du, 2023). Building upon this, we encourage models to reflect, revise, and eliminate, to enhance their reasoning reliability.

2.3. Narrative Framework

Strategies that emphasize narrative flow, event structure, and thematic coherence have been shown to improve comprehension and support commonsense reasoning. The National Institute of Child Health and Human Development reviewed 203 studies on text comprehension instruction and identified two effective instructional strategies: teaching readers to answer "who, what, where, when, and why" questions about a narrative, and instructing readers to focus on the story structure by posing similar questions to themselves as they read (Eunice Kennedy Shriver, 2000). These methods help readers understand the causal and thematic coherence of a story.

Recent work demonstrates that storytelling serves as an effective inductive bias for improving LLMs' performance on reasoning tasks involving temporal structure. The Narrative-of-Thought (NoT) framework shows that guiding models to construct temporally grounded narratives from unordered event sets before predicting their temporal order significantly boosts performance on temporal graph generation tasks, underscoring the value of narrative-based prompting for reasoning tasks. More specifically, rather than directly prompting the model to predict the correct ordering of events, NoT first asks the model to generate a coherent story connecting the events, and then construct the structured temporal graph (Zhang et al., 2024).

Inspired by this approach, we apply narrative-based prompts to HellaSwag, using narrative comprehension techniques to give LLMs cognitive scaffolds to improve commonsense reasoning accuracy.

2.4. Linguistic Cues Frameworks

Among various cognitive prompting strategies, we focus particularly on linguistic cues. Reading comprehension has been defined by psychologist Dolores Durkin as "intentional thinking during which meaning is constructed through interactions between text and reader" (1993). This process depends heavily on the reader's ability to interpret linguistic cues (Durkin, 1993; National Reading Panel, 2000), such as transition words and prepositions, which signal temporal, spatial, and causal relationships essential for commonsense reasoning.

For example, when a reader encounters the word *however*, they may pause to reassess previously stated information, demonstrating an awareness of contrast. Likewise, prepositions such as *behind*, *next to*, or *above* establish essential spatial relationships.

LLMs appear to exhibit similar sensitivities to linguistic cues. For instance, Light et al. (2025) demonstrate that function words, such as *which* and *therefore*, significantly influence an LLM's interpretations. Similarly, Zhang et al. (2019) show that modifying the model's internal architecture to attend to key linguistic elements improves performance. While such work highlights the role of linguistic cues, it largely centers on architectural changes. In contrast, prompt-based methods for directing attention to linguistic cues, like transition words and prepositions, remain largely underexplored for improving model reasoning.

3. Methods

3.1. Experimental Design

To evaluate the performance of our cognitively-based prompting strategies, we conduct experiments using the 2019 HellaSwag dataset (Zellers et al., 2019) and the GPT-3.5 Turbo model. Specifically, the dataset tests an LLM's ability to complete sentences depicting short scenarios with one of four multiple-choice sentence endings, only one of which is logically coherent. Given its emphasis on commonsense reasoning in everyday contexts, HellaSwag serves as an appropriate benchmark for assessing psychologically-grounded reasoning frameworks. We use a uniformly randomized subset of 400 examples from the HellaSwag dataset to ensure manageability while maintaining diversity in scenario types.

Our prompting strategies fall into four psychology-inspired framework categories: metacognitive-based, strategy-based, narrative-based, linguistic cue-based reasoning. Within each category, we evaluate specific frameworks (detailed in the next section). Additionally, we evaluate the impact of few-shot vs. zero-shot prompting. Specifically, we include example questions with the correct answer and detail an example of going through the reasoning steps for the prompting strategy (see Appendix C).

For baseline comparisons, we establish a lower bound using a generic prompt in GPT-3.5 Turbo. Upper bounds are estimated using the GPT-o4-mini reasoning model and a more powerful GPT-4.1 model. We also evaluate performance against established reasoning strategies, including CoT prompting and CoT with few-shot prompting. This allows us to benchmark how much cognitive-inspired prompting strategies can close the performance gap between model versions.

3.2. Prompting Frameworks

3.2.1. Metacognitive Reasoning Frameworks

METAL Reasoning Framework METAL is a metacognition-inspired strategy designed to focus on problem comprehension, deliberate reasoning, counter-arguments, and self-evaluation. Each letter in METAL corresponds to a distinct phase of reasoning:

- **M:** Make a claim
- **E:** Explain your reasoning
- **T:** Think of a tempting alternative
- **A:** Acknowledge limits
- **L:** Learn from it

It builds upon the stages in the Metacognitive Prompting (MP) framework (Wang & Zhao, 2024), specifically by including the "T" and "A" stages that relate to counterargument construction and self-critique, cognitive strategies known to enhance reasoning skills (Kuhn & Udell, 2003). In addition, the "L" stage promotes reflective learning.

RNRRR Reasoning Framework RNRRR is an iterative self-refinement strategy inspired by metacognition, the Self-Regulated Learning (SRL) model (Zimmerman, 2002), and reflective problem-solving theories (Flower & Hayes, 1981; Simon, 2012; Madaan et al., 2023).

It integrates structured reflection and narrative coherence into a reasoning process that unfolds across five key stages:

- **R:** Read and choose
- **N:** Narrate the fit
- **R:** Rival and reject
- **R:** Reflect on risks
- **R:** Reaffirm with strength

The five stages of RNRRR correspond to key components of the SRL model of information intake, strategic planning, reflecting, and revising. The "N" draws subtle inspiration from narrative structures, engaging cognitive storytelling mechanisms.

3.2.2. Strategic Reasoning Framework

POELO Reasoning Framework This strategy draws from the POE framework (Ma & Du, 2023) as options are ruled out and the most plausible choice is selected from the remaining candidates:

- **P:** Preview and pick
- **O:** Outrule one
- **E:** Examine another
- **L:** Look again
- **O:** Opt with confidence

The strategy follows strategic elimination and reflective reassessment.

3.2.3. Narrative Frameworks

Thematic Flow Reasoning Thematic Flow Reasoning is a narrative-based prompting strategy designed to help LLMs identify realistic story continuations by tracking emotional and thematic coherence. It draws on cognitive and

educational psychology findings, which show that readers improve their comprehension when guided to focus on narrative elements such as tone, activity type, and underlying themes (Eunice Kennedy Shriver, 2000).

Step 1: Situation Assessment

- The emotional tone of the scene
- The type of activity being described
- The overarching theme

Step 2: Option Evaluation

- Whether it maintains the emotional tone
- Whether it is thematically and situationally consistent
- Whether it avoids abrupt or implausible shifts

5W Prompting 5W Prompting builds on journalistic and educational techniques that encourage detailed situational understanding by answering the questions Who, What, Where, When, and Why. This framework aligns with the story structure, which helps students evaluate coherence across characters, actions, space, time and intent in story comprehension (Eunice Kennedy Shriver, 2000).

Step 1: Narrative Grounding

- Who is involved
- What is happening
- Where the scene is located
- When it occurs in the sequence
- Why the action is happening

Step 2: Option Comparison

- Consistency in the answers to who, what, where, when, and why
- Whether any option introduces unjustified shifts or contradictions

3.2.4. Linguistic Cue Frameworks

The ability to interpret linguistic cues is critical for effective reasoning (Durkin, 1993; National Reading Panel, 2000). Cues such as transition words and prepositions signal temporal, spatial, and causal relationships that support common sense understanding. To target these cues, we designed three prompt templates, each based on a different type of linguistic signal:

Transition Words: Prompts included the top 10 most common transition words that emphasize temporal order and cause-and-effect relationships, such as *First*, *Next*, *However*, and *Because*.

Prepositions: Prompts focused on the top 10 most common prepositions that show spatial or temporal relationships, like *Behind*, *Next to*, *During*, and *By*.

Transitions and Prepositions: Prompts included both the transition words and prepositions together to test if using both improves model reasoning.

Examples of the exact prompts used are provided in Appendix B.

3.3. Evaluation Metrics

To evaluate the reliability of different prompting strategies, we use two complementary metrics: accuracy and consistency. Together, these metrics provide a comprehensive view of reliability, capturing both how often a strategy succeeds and how consistently it does so.

Accuracy is measured as the proportion of correctly answered questions, and we report statistical significance using p-values to assess whether the strategies meaningfully improve over the baseline. We show via permutation tests that the results are significant at the 5% level, indicating most strategies are effective in improving commonsense reasoning.

Consistency is assessed through the standard deviation of accuracy scores across trials. A lower standard deviation indicates more stable performance, implying that the strategy produces reliable results across runs.

Both consistency and accuracy are critical for evaluating reliability and robustness.

4. Results and Analysis

4.1. Evaluating Accuracy

We evaluated the performance of various cognitive-based prompting strategies across zero-shot and few-shot settings, comparing them to multiple baselines. All proposed zero-shot reasoning strategies show an increase in accuracy, as seen in Figure 1. At a 5% significance level for the permutation test, all zero-shot strategies, except 5Ws and transition words, outperform the GPT-3.5 Turbo baseline as seen in Table 1. These findings suggest that prompting strategies grounded in psychology have potential towards excelling in tasks requiring complex or commonsense reasoning.

Moreover, the zero-shot versions of METAL, thematic flow, and POELO statistically outperform CoT zero-shot benchmarks, additionally demonstrating how cognitive-based

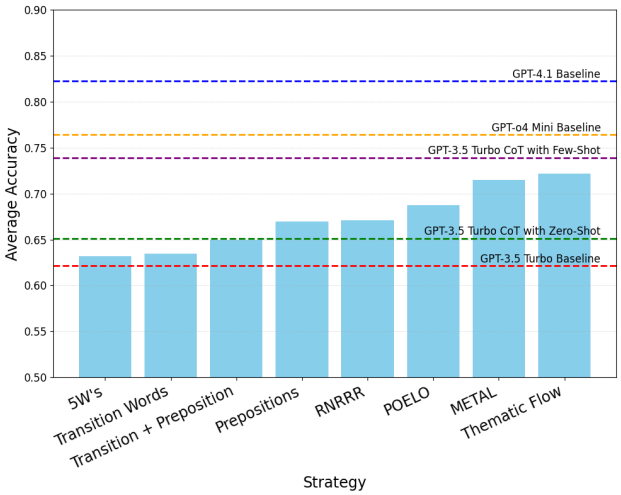

Figure 1. Average accuracy of cognitive-based strategies with zero-shot prompting in ascending order. Dashed lines represent benchmark accuracies.

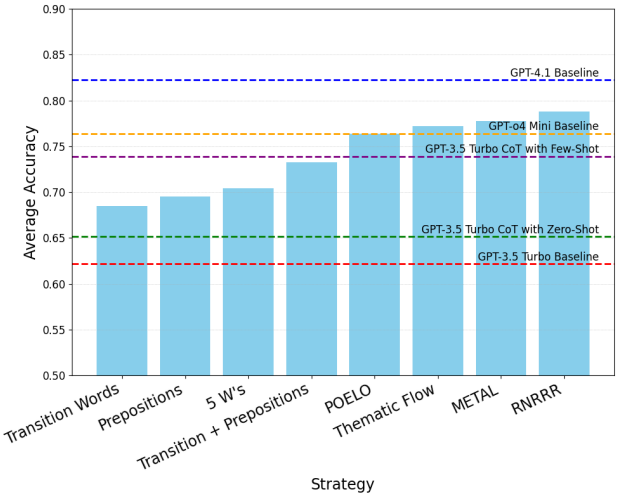

Figure 2. Average accuracy of cognitive-based strategies with few-shot prompting in ascending order. Dashed lines represent benchmark accuracies.

Table 1. Permutation test p-values for all prompting strategies relative to selected evaluation benchmarks. Lower values indicate stronger evidence of significance. The strategies are METAL, thematic flow (TF), transition words, prepositions, transitions and prepositions (T & P), RNRRR, POELO, and 5W.

Prompting Strategy	**METAL**	**TF**	**Transition**	**Prepositions**	**T & P**	**RNRRR**	**POELO**	**5W**
Zero-shot (vs. Baseline)	0.0040	0.0040	0.1587	0.0040	0.0159	0.0079	0.0040	0.1349
Few-shot (vs. Baseline)	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040
Zero-shot (vs. CoT Zero-shot)	0.0040	0.0040	0.8770	0.0992	0.5675	0.0675	0.0198	0.9365
Few-shot (vs. CoT Zero-shot)	0.0040	0.0040	0.0238	0.0119	0.0040	0.0040	0.0040	0.0079
Few-shot (vs. CoT Few-shot)	0.0040	0.0040	1.0000	1.0000	0.7738	0.0040	0.0238	1.0000

frameworks can outperform established reasoning strategies. While thematic flow and METAL stand out in zero-shot settings, they do not surpass the CoT few-shot benchmark, likely due to the absence of explicit examples that show the intended output.

All proposed few-shot reasoning strategies show statistically significant accuracy improvements over both the GPT-3.5 Turbo baseline and the zero-shot CoT benchmark, as confirmed by permutation testing at the 5% significance level as illustrated in Table 1 and Figure 2. This aligns with expectations given the advantages of few-shot prompting.

Additionally, RNRRR, METAL, POELO and thematic flow, beat the CoT with few-shot benchmark at a 5% significance level. Of these, all but POELO also beat the reasoning GPT-o4-mini model, showing that metacognitive-based prompting and narrative-based prompting perform best. It can allow GPT-3.5 Turbo to outperform established reasoning models and other established reasoning prompting methods. These cognitive-based frameworks allow GPT-3.5 Turbo to perform better than GPT-o4-mini reasoning.

We also observe a broader trend where few-shot variants of cognitive-inspired strategies consistently outperform their zero-shot counterparts, underscoring the benefits of combining cognitive frameworks with example-based prompting. Metacognitive and narrative strategies in particular, such as RNRRR's self-regulation loop, METAL's integration of metacognitive and strategic reasoning, and Thematic Flow's emphasis on coherence, achieve higher accuracy than linguistically-oriented strategies, reinforcing their value for reliable language model reasoning.

See Appendix A for additional results.

4.2. Evaluating Consistency

We evaluate consistency using the standard deviation of accuracy scores. Lower standard deviation indicates reduced variability and more reliable performance across runs. While a model may perform well on average, high variability can undermine trust, especially in high-stakes contexts.

As illustrated in Figure 3, all cognitively-inspired strategies

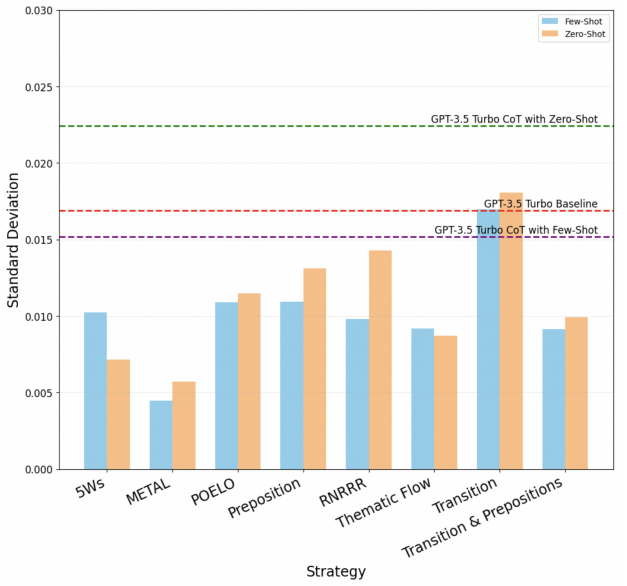

Figure 3. Standard deviation of accuracy scores few-shot (blue) and zero-shot (orange) cognitive-based strategies. Dashed lines represent benchmark standard deviations.

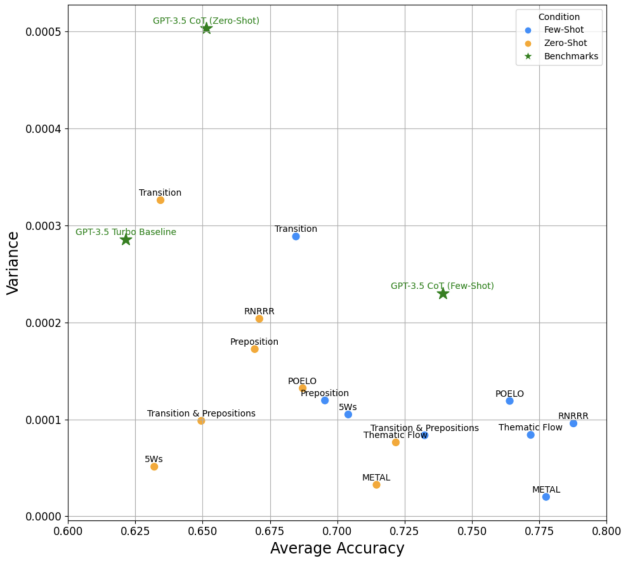

Figure 4. Average accuracy versus variance of prompting strategies under few-shot (blue) and zero-shot (orange) conditions. Benchmarks are shown as green stars.

have lower standard deviation than GPT-3.5 Turbo CoT with zero-shot prompting. With the exception of the Transition Words framework, all strategies also achieve greater consistency than GPT-3.5 baseline prompt and few-shot CoT benchmark.

The higher standard deviation observed in the Transition Words framework may reflect limitations in the clarity of the linguistic signals provided. The METAL framework stands out as particularly stable, achieving the lowest standard deviation across few-shot and zero-shot prompting models. One explanation is that structured metacognition reduces variability, as think-aloud and evaluation steps anchor reasoning, and implicit error correction occurs.

4.3. Evaluating Accuracy and Consistency Together

While accuracy and consistency were analyzed separately previously, examining them together provides a more comprehensive assessment of reliability.

We see that few-shot variants of METAL, Thematic Flow, RNRRR, and POELO consistently outperform GPT-3.5 benchmarks, achieving both higher accuracy and consistency as seen in Figure 4. METAL stands out for its exceptional consistency, especially in few-shot settings. Although few-shot RNRRR's variance is slightly higher than few-shot METAL's variance, its accuracy gain is nearly quadruple that margin. This trade-off suggests that the improvement in accuracy outweighs the reduction in consistency, positioning RNRRR with few-shot as the most effective overall cognitive-inspired strategy we tested.

Together, these findings highlight that cognitively grounded prompting strategies not only improve performance but also enhance consistency, reinforcing their value for reliable language model reasoning.

5. Conclusion

Our findings underscore that cognitive-based prompting strategies offer a powerful way to enhance the reliability of LLMs by improving both the accuracy and consistency of their responses in complex logical and commonsense reasoning tasks. Notably, metacognitive-based prompting frameworks like METAL and RNRRR boost model performance and show consistency across responses, which is a key metric for dependable behavior. Our study was constrained by the available resources, which limited the scope of our analysis to a subset of the HellaSwag dataset. This work opens several avenues for extension. One direction is evaluating whether the proposed prompting strategies generalize across a wider range of reasoning benchmarks. Another is exploring combinations of prompting frameworks, including integrating our approach with established strategies such as retrieval-augmented generation and self-consistency. Additionally, future research could examine multi-agent paradigms. Ultimately, cognitive-based prompting strategies offer a promising path toward developing AI systems that are not only more capable but also more reliable and

aligned with principles of responsible AI.

Impact Statement

This work aims to advance the field of Machine Learning by introducing structured, cognitively-inspired prompting strategies that enhance the overall reliability of LLMs. By enhancing performance on commonsense reasoning tasks and promoting stable outputs across trials, these strategies become especially valuable in high-stakes domains such as education, law, and clinical decision support, where accuracy and consistency are essential. In this way, our approach helps mitigate key risks in foundation models, such as hallucinations. Additionally, the structured nature of these prompts improves the interpretability of model outputs, contributing to the development of more reliable, robust, and responsible language models.

However, as with any prompting method, the potential for hallucinations and incorrect reasoning remains. It is therefore critical that users remain aware of these limitations. We encourage continued exploration and refinement of cognitive-inspired prompting strategies, with an emphasis on improving their reliability, transparency, and integration into broader frameworks for the safe and responsible deployment of language models in real-world applications.

6. Acknowledgments

We thank Professor Yisong Yue, Alexander Farhang, Hao Liu, and Geeling Chau for their valuable discussions and guidance throughout the development of this project. We also acknowledge the CS 159 course for funding API usage.

References

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. 2020. doi: 10.48550/ARXIV.2005.14165. URL https://arxiv.org/abs/2005.14165.

Durkin, D. *Teaching Children To Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Allyn and Bacon, Boston, 6th ed. edition, 1993. ISBN 9780205139156.

Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. Report of the national reading panel: Teaching children to read: Reports of the subgroups (reference only). Technical Report 00-4754, U.S. Government Printing Office, Washington, DC, 2000. NICHD Publication No. 00-4754.

Flavell, J. H. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906–911, 1979. doi: 10.1037/0003-066X.34.10.906. URL https://doi.apa.org/doi/10.1037/0003-066X.34.10.906.

Flower, L. and Hayes, J. R. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365, 1981. doi: 10.2307/356600. URL https://www.jstor.org/stable/356600?origin=crossref.

Kramer, O. and Baumann, J. Unlocking structured thinking in language models with cognitive prompting. (arXiv:2410.02953), November 2024. doi: 10.48550/arXiv.2410.02953. URL http://arxiv.org/abs/2410.02953. arXiv:2410.02953.

Kuhn, D. and Udell, W. The development of argument skills. *Child Development*, 74(5):1245–1260, 2003. doi: 10.1111/1467-8624.00605. URL https://srcd.onlinelibrary.wiley.com/doi/10.1111/1467-8624.00605.

Li, Z., Peng, B., He, P., Galley, M., Gao, J., and Yan, X. Guiding large language models via directional stimulus prompting, 2023. URL https://arxiv.org/abs/2302.11520.

Light, J., Cheng, W., Yue, W., Oyamada, M., Wang, M., Paternain, S., and Chen, H. Disc: Dynamic decomposition improves llm inference scaling, 2025. URL https://arxiv.org/abs/2502.16706.

Liu, F., Emerson, G., and Collier, N. Visual spatial reasoning, 2022. URL https://arxiv.org/abs/2205.00363.

Ma, C. and Du, X. Poe: Process of elimination for multiple choice reasoning. (arXiv:2310.15575), October 2023. doi: 10.48550/arXiv.2310.15575. URL http://arxiv.org/abs/2310.15575. arXiv:2310.15575.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. (arXiv:2303.17651), May 2023. doi: 10.48550/arXiv.2303.17651. URL http://arxiv.org/abs/2303.17651. arXiv:2303.17651.

Simon, H. A. *The Architecture of Complexity*, pp. 335–361. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 9783642279218 9783642279225. doi: 10.1007/978-3-642-27922-5_23. URL http://link.springer.com/10.1007/978-3-642-27922-5_23.

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. (arXiv:2307.09288), July 2023. doi: 10.48550/arXiv.2307.09288. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288.

Wang, Y. and Zhao, Y. Metacognitive prompting improves understanding in large language models. (arXiv:2308.05342), March 2024. doi: 10.48550/arXiv.2308.05342. URL http://arxiv.org/abs/2308.05342. arXiv:2308.05342.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. 2022. doi: 10.48550/ARXIV.2201.11903. URL https://arxiv.org/abs/2201.11903.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? 2019. doi: 10.48550/ARXIV.1905.07830. URL https://arxiv.org/abs/1905.07830.

Zhang, X. F., Beauchamp, N., and Wang, L. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. (arXiv:2410.05558), November 2024. doi: 10.48550/arXiv.2410.05558. URL http://arxiv.org/abs/2410.05558. arXiv:2410.05558.

Zhang, Z., Wu, Y., Zhou, J., Duan, S., Zhao, H., and Wang, R. Sg-net: Syntax-guided machine reading comprehension, 2019. URL https://arxiv.org/abs/1908.05147.

Zimmerman, B. J. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, 2002. doi: 10.1207/s15430421tip4102_2. URL http://www.tandfonline.com/doi/abs/10.1207/s15430421tip4102_2.

Zohar, A. and David, A. B. Explicit teaching of meta-strategic knowledge in authentic classroom situations. *Metacognition and Learning*, 3 (1):59–82, 2008. doi: 10.1007/s11409-007-9019-4. URL http://link.springer.com/10.1007/s11409-007-9019-4.

A. Additional Results

Table 2. P-values and observed differences for prompting strategies relative to different evaluation baselines. Lower p-values indicate stronger statistical evidence that performance differs from the compared baseline. The strategies are METAL, thematic flow (TF), transition words, prepositions, transitions and prepositions (T & P), RNRRR, POELO, and 5W.

Prompting Strategy	**METAL**	**TF**	**Transition**	**Prepositions**	**T & P**	**RNRRR**	**POELO**	**5W**
P-values								
Zero-shot (vs. Baseline)	0.0040	0.0040	0.1587	0.0040	0.0159	0.0079	0.0040	0.1349
Few-shot (vs. Baseline)	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040
Zero-shot (vs. CoT Zero-shot)	0.0040	0.0040	0.8770	0.0992	0.5675	0.0675	0.0198	0.9365
Few-shot (vs. CoT Zero-shot)	0.0040	0.0040	0.0238	0.0119	0.0040	0.0040	0.0040	0.0079
Zero-shot (vs. CoT Few-shot)	0.9921	0.9603	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Few-shot (vs. CoT Few-shot)	0.0040	0.0040	1.0000	1.0000	0.7738	0.0040	0.0238	1.0000
Observed differences								
Zero-shot (vs. Baseline)	0.0931	0.1002	0.0129	0.0478	0.0280	0.0524	0.0656	0.0105
Few-shot (vs. Baseline)	0.1560	0.1503	0.0631	0.0739	0.1109	0.1658	0.1425	0.0826
Zero-shot (vs. CoT Zero-shot)	0.0631	0.0703	-0.0170	0.0180	-0.0019	0.0225	0.0357	-0.0195
Few-shot (vs. CoT Zero-shot)	0.1261	0.1204	0.0332	0.0440	0.0810	0.1360	0.1126	0.0527
Zero-shot (vs. CoT Few-shot)	-0.0246	-0.0174	-0.1048	-0.0698	-0.0896	-0.0652	-0.0520	-0.1071
Few-shot (vs. CoT Few-shot)	0.0384	0.0327	-0.0545	-0.0437	-0.0067	0.0482	0.0249	-0.0351

B. Specific Prompts Used

```
You are a thoughtful reasoner using the METAL framework:

M: Make a claim — state which answer you think is correct.
E: Provide evidence — explain why that answer fits best.
T: Think of a counter — consider another choice that might seem plausible.
A: Acknowledge limitations — note any uncertainty or assumptions.
L: Learn from it — reflect on what makes the chosen answer better overall.

Given a multiple-choice question (MCQ), select the number (0, 1, 2, or 3)
corresponding to the correct answer.

Do not output the answer until all 5 steps are completed.
End with: ANSWER: number
```

Figure 5. METAL prompt example.

```
R: Read and choose — Pick the option (0-3) that best continues the scenario.
N: Narrate the fit — Explain how your choice naturally follows from the story so far.
R: Rival and reject — Identify one tempting but incorrect option, and explain why it
fails.
R: Reflect on risks — Acknowledge ambiguity, tone shifts, or missing details that
could make your answer uncertain.
R: Reaffirm with strength — Justify why your original answer still stands out as the
best continuation.

Only after completing all 5 steps, end with: ANSWER: <number>
```

Figure 6. RNRRR prompt example.

```
P: Preview and pick — Read all four options and choose the one that initially seems
most plausible.
O: Outrule one — Identify one option that clearly doesn't fit based on logic, tone, or
continuity, and eliminate it.
E: Examine another — Consider a second option that feels tempting. Explain why it
ultimately falls short.
L: Look again — Reassess your original pick in light of the eliminated choices. Does
it still hold up?
O: Opt with confidence — Justify your final choice clearly and decisively.

Only after completing all 5 steps, end with: ANSWER: <number>
```

Figure 7. POELO prompt example.

```
Given the question and choices, choose the number (0, 1, 2, or 3) corresponding to the
best continuation of the underlying emotional and thematic flow.

    Step 1:
    When reading the question, focus specifically on understanding the following:
    1. What is the emotional tone? (for example, happy, tense, casual, tragic)?
    2. What is the type of activity? (for example, socializing, competing, relaxing,
problem-solving?)
    3. What is the bigger theme? (for example, friendship, fun, work stress, danger?)

    Step 2:
    Then for each of the multiple choice options 0, 1, 2, or 3, consider:
    1. Does it continue the emotional tone?
    2. Does it fit the same kind of theme or activity? Or, does it feel abrupt,
mismatched, or out of the emotional rhythm?

    Choose the most realistic and coherent continuation — it should follow naturally
from the scene and stay in the same context.
    Avoid answers that introduce unrelated actions or illogical jumps. When unsure,
prefer the option that maintains the underlying emotional and thematic flow.
```

Figure 8. Thematic Flow prompt example.

```
Given the question and four options, identify the most coherent continuation by
answering the following:


    - WHO is involved? (Identify the main subjects — person, animal, object)
    - WHAT is happening? (Summarize the current action or event)
    - WHERE is it happening? (Identify the physical setting)
    - WHEN is this happening? (Time of day, phase of activity, sequence)
    - WHY is it happening? (Purpose or intention — implied or stated)


    For each option:
    - Does it involve the same WHO?
    - Is the WHAT a natural continuation of the earlier action?
    - Does the WHERE stay consistent?
    - Does the WHEN make sense in the sequence?
    - Is the WHY reasonable, or does it contradict the scene?


    Eliminate options that introduce new characters, places, or implausible motives.
```

Figure 9. 5W prompt example.

```
Given the mcq, choose 0, 1, 2, or 3 as the number corresponding to the correct answer.
Pay attention to key transition words when working through the logic, like
"First," "Then," "Next," "However," "Although," "Furthermore," "Also," "Because," etc.
End your answer with: ANSWER: number
```

Figure 10. Transition words prompt example.

```
Given the mcq, choose 0, 1, 2, or 3 as the number corresponding to the correct answer.
Pay attention to key prepositions when working through the logic which can help
understand spatial relationships, like "Under," "Behind," "Next to," "Beside,"
"At," "From," "During," "By," etc.
End your answer with: ANSWER: number
```

Figure 11. Prepositions prompt example.

```
Given the question and choices, choose the number (0, 1, 2, or 3) corresponding to the
best continuation.

Focus on:
- Transition words (e.g., "Then," "Next," "After") — they signal what should logically
happen next.
- Prepositions (e.g., "on," "under," "behind") — they help with spatial and temporal
grounding.

The best choice should feel like a realistic and smooth next step, following naturally
in both time and space.

Output format:
ANSWER: number
```

Figure 12. Transition and Prepositions prompt example.

C. Few-shot Prompting Example

```
Example 1:
Question: Then, the man writes over the snow covering the window of a car, and a woman
wearing winter clothes smiles. then
Options:
0: the man adds wax to the windshield and cuts it.
1: a person boards a ski lift, while two men supporting the head of the person wearing
winter clothes snow as the we girls sled.
2: the man puts on a Christmas coat, knitted with netting.
3: the man continues removing the snow on his car.

M: I claim that option 3 is correct.
E: The man is already interacting with snow on a car, so continuing to remove it is a
logical next step.
T: Option 2 might seem plausible since it's winter and he might put on winter gear.
A: However, there's no prior mention of needing new clothes or a costume change — and
option 2 lacks relevance.
L: The clearest progression is from writing in snow to removing it, making 3 the most
natural continuation.
ANSWER: 3
```

Figure 13. Example of a demonstration question included in the METAL prompt to guide few-shot LLM reasoning.

D. Example Response

```
Question: Several food items and dishes are laid out on a table. Meat product and
other items are used to create a sandwich. then a bento shaper
0: is put into a skillet on the stove.
1: is used to create an image in the sandwich.
2: is put together and put into the oven.
3: depositions and improvement of the sandwich.
```

```
5W Reasoning:
- WHO: Not explicitly mentioned, but likely a person preparing food.
- WHAT: Making a sandwich using different food items.
- WHERE: In a kitchen setting.
- WHEN: During the preparation of the sandwich.
- WHY: Likely for consumption or presentation.

Evaluate options:
Option 0: Involves a skillet and stove - introduces new elements not previously
mentioned. Inconsistent with making a sandwich.
Option 1: Creates an image in the sandwich - a creative, plausible step in sandwich
preparation.
Option 2: Puts the sandwich in the oven - a common practice in cooking, fits the
context of food preparation.
Option 3: Unclear language ("depositions and improvement of the sandwich") - doesn't
fit the context of making a sandwich.

The most coherent continuation is making an image in the sandwich.
ANSWER: 1
```

Figure 14. Example of a prompt given to the LLM and its 5W reasoning-based response.