

Beyond “Think Step-by-Step”: Enhancing Abstract Analogy Reasoning in Medium-Sized LLMs with Structured CoT

Jakub Podolak Önder Akaçık Srishti Yadav
Mina Janićijević Dylan Goode Milan Miletić

University of Amsterdam

1 Introduction

Large language models (LLMs) such as LLaMA-3, Starling, or Mistral are commonly evaluated in few-shot or zero-shot settings by prompting them with task-specific instructions and/or demonstration examples. A key concept in improving the performance of these models is Chain-of-Thought (CoT) prompting, where intermediate reasoning steps are produced until the LLM arrives at its final prediction (Wei et al., 2022; Kojima et al., 2022; Suzgun et al., 2022). This has been particularly effective for complex reasoning tasks.

Another such complex task is that of abstract analogy reasoning, which is considered central to human intelligence, as it enables structured comparisons to solve unfamiliar problems and fosters creativity and scientific innovation. However, this task still poses a considerable challenge even for very large LLMs (Webb et al., 2023).

The main contribution of this research is three-fold:

1. We evaluate the effectiveness of zero-shot CoT prompting (Kojima et al., 2022) in medium-sized LLMs, specifically LLaMA-3 8B, Starling 7B, and Mistral 7B, for the task of abstract analogy resolution
2. We experiment with different prompt variants and apply them on two separate abstract analogy tasks (story analogy and verbal analogy), in order to verify the robustness of this approach
3. We introduce a novel approach to zero-shot CoT that incorporates task-specific steps. We call this approach Structured CoT.

Previous research has demonstrated the efficacy of CoT prompting in large LLMs (Wei et al., 2022; Suzgun et al., 2022), but recent advancements in the field motivated us to re-assess its impact on newly available medium-sized LLMs that

have been trained using novel methods, such as instruction-tuning, RLHF or RLAI. Furthermore, while studies have addressed the task of analogy resolution in LLMs (Webb et al., 2023; Hofstadter and Mitchell, 1994), to the best of our knowledge there has not been an extensive study of the impact of CoT for this particular task.

Finally, in addition to the traditional zero-shot CoT prompting approach, in which "Let's think step-by-step" is appended to the prompt, we introduce structured CoT prompting. This method adds a set of task-specific instructions to the prompt designed not only to elicit CoT reasoning but also to guide the reasoning steps in a structured way relevant to the task.

Our findings indicate that standard zero-shot CoT prompting improves performance on the verbal analogy task but fails to do so on the task of story analogy. In contrast, structured CoT prompting improves performance on story analogy task for two out of three models and matches the performance of standard zero-shot CoT on verbal analogy task. Moreover, the models generally exhibit robustness to different prompt variants, but benefit from incorporating these variants together. These results underscore the potential of both traditional and structured CoT prompting in enhancing the reasoning capabilities of medium-sized LLMs. The code for the project can be seen [here](#).¹

2 Related Work

Chain-of-Thought (CoT) prompting has emerged as a powerful technique to enhance the reasoning capabilities of large language models (LLMs). Wei et al. (2022) introduced CoT prompting, which generate intermediate reasoning steps to improve the performance of LLMs on complex tasks such as arithmetic, commonsense, and symbolic reasoning. Their experiments with models like PaLM, GPT-3,

¹<https://github.com/jakub-podolak/AbstractAnalogies>

and LaMDA showed that CoT prompting significantly boosts performance, particularly in tasks requiring multi-step reasoning. They found that benefits of CoT prompting are more pronounced in models with approximately 100B parameters or more.

Building on the foundation laid by Wei et al. (2022), Kojima et al. (2022) explored the potential of zero-shot CoT prompting. This method employs simple phrases such as "Let's think step by step" to elicit reasoning capabilities in LLMs even without few-shot examples. Their findings revealed substantial improvements across various reasoning benchmarks, such as arithmetic and logical tasks. Zero-shot CoT demonstrated versatility and task-agnostic effectiveness, suggesting that inherent zero-shot reasoning capabilities can be unlocked with minimal prompting.

Webb et al. (2023) addressed the challenge of analogical reasoning, a critical aspect of human intelligence, by evaluating the performance of GPT-3 on a range of analogy tasks. They found that GPT-3, even without direct training, exhibited a strong capacity for capturing abstract patterns, often surpassing human² performance (e.g., verbal analogy task). However, GPT-3's performance was less robust in tasks requiring more abstract causal relations (e.g., story analogy task).

Despite the demonstrated effectiveness of CoT prompting in models with more than 100B parameters, significant advancements have been made in LLMs since the papers by Wei et al. (2022) and Kojima et al. (2022) were published. Evaluating new models, such as LLaMa-3 8B³, Mistral 7B (Jiang et al., 2023), and Starling 7B-beta (Zhu et al., 2023) is worthwhile to revisit the effectiveness of CoT prompting in medium-sized LLMs

Our study builds on the foundational work of Wei et al. (2022) and Kojima et al. (2022) by applying CoT prompting to medium-sized LLMs and examining its impact on a subset of abstract analogy reasoning tasks introduced in Webb et al. (2023). By comparing performance across different model sizes and prompt types, we seek to understand how CoT prompting influences medium-sized LLMs reasoning capabilities and whether these models can approximate the performance of their larger counterparts.

3 Methods

This section describes the methodology used in the project. It includes details on the selection of models, preparation of prompts, and the evaluation metrics used to assess model performance.

3.1 Model Selection

In selecting the models for our study, we focused on medium sized LLMs known for their effectiveness in various NLP tasks despite having fewer parameters compared to the larger frontier models. LLaMa-3 8B⁴ specifically the instruction-tuned version, Starling 7B (Zhu et al., 2023), and Mistral 7B, also in its instruction-tuned form (Jiang et al., 2023). Starling 7B was selected for its proven capabilities in complex reasoning tasks essential for testing CoT prompting (Zhu et al., 2023). The instruction-tuned versions of LLaMA-3 and Mistral are particularly suited for our study as they are optimized for processing and responding to natural language instructions (Gupta et al., 2023).

3.2 Task & Datasets

In this section, we describe the tasks (verbal analogies and story analogies) and datasets used to evaluate the performance of aforementioned models on abstract analogy reasoning.

3.2.1 Verbal Analogies

Following the work of Lu et al. (Lu et al., 2019), we used a dataset called UCLA Verbal Analogy Dataset⁵, comprising of four-term verbal analogy problems involving a broader range of semantic relations. The dataset contains a series of four-term analogy problems in the form 'A:B::C:?' which we used to create analogy task in the form of "A is to B as C is to ?". This involves presenting a relationship between two words (A is to B) and a new word (C), where the task is to determine which word out of two choices, D and D', is analogous to the given word pair (A is to B). This particular dataset includes four types of analogies: synonym, opposite, function, and category members (details provided in the Appendix).

Verbal analogies were chosen because they are a common task for assessing the models' ability to capture analogical reasoning in terms of recognizing abstract word relationships. While this task is generally considered easier than story analogies, it

²Fifty-seven UCLA undergraduates

³<https://llama.meta.com/llama3/>

⁴<https://llama.meta.com/llama3/>

⁵UCLA Verbal Analogy Dataset

serves as a fundamental benchmark for evaluating analogical reasoning capabilities in LLMs.

Verbal Analogy	
Synonym	easy : simple :: sad : ?
Opposite	love : hate :: rich : ?
Function	burn : fire :: blow : ?
Category Members	bird : crow :: sport : ?

Table 1: **Example samples from the dataset for verbal analogies** (synonym, opposite, function, and category members).

3.2.2 Story Analogies

Furthermore, we evaluated the models on a set of far story analogies derived from (Gentner et al., 1993). This dataset includes 18 source stories, each accompanied by two potential target stories. Both target stories are matched with the source story in terms of first-order relations, but only one shares the same causal relations as the source. The task involves identifying which of the two target stories is more analogous to the source story based on these higher-order causal relations.

Story analogies were chosen because they represent a more complex and challenging form of analogy resolution, as they go beyond individual concepts such as words. Higher-order relations, particularly causal relations between events, play a central role in human analogical reasoning and are crucial for understanding complex real-world events. In the study by Webb et al. (2023), story analogy was the only task where models performed worse than human participants, highlighting the difficulty and importance of this task. Focusing on far analogies, which involve the same first-order relations as the source story, but use different entities, provides a rigorous test of the models’ ability to recognize and apply higher-order relations across different contexts.

3.3 Prompt Preparation

To test whether CoT prompting and model reasoning improve performance in both story and verbal analogies, we prepared four different types of prompts. These prompts are designed to evaluate the models’ reasoning capabilities under various conditions. The four prompt types are:

- **Basic Prompt:** The model is given the task setup and asked for the answer. This prompt simply presents the problem and the target options, asking the model to identify the correct analogy.
- **Basic Prompt Forced:** The basic prompt is modified to enforce the model to provide only the final answer, turning off any reasoning. This decision is motivated by the fact that newer models are often trained to provide some reasoning by default, and this prompt type ensures that no intermediate reasoning steps are generated.
- **Zero-shot CoT:** The basic prompt is extended with the phrase "Let’s think step by step." This prompt aims to encourage the model to generate intermediate reasoning steps before arriving at the final answer.
- **Structured CoT:** The basic prompt is supplemented with a set of task-specific instructions, providing a structured path of reasoning. This prompt was manually prepared by the authors to guide the model through a series of reasoning steps relevant to the task.

4 Experiments

4.1 Model Settings

As mentioned in the methods section, we employed medium-sized language models LLaMA-3 8B, Starling 7B, and Mistral 7B to evaluate their performance on analogy resolution tasks. To ensure consistency and comparability across all tests, we configured each model with specific settings using the Hugging Face Transformers library. The tokenizer employed was AutoTokenizer, which is standard for these models. We set the maximum number of new tokens generated (max_new_tokens) to 2048 to allow sufficient room for the models to generate complete responses. Sampling was disabled (do_sample set to False) to use greedy encoding, ensuring that the models always chose the most likely next token. The temperature parameter was not applicable in our setup due to the use of greedy encoding, which does not require stochasticity in token selection.

4.2 Pipeline

Majority Vote To prevent bias from the prompt design itself, we prepared three slightly different

Task	Basic Prompt Example
Story Analogy	<p>This is a source story: {SourceStory}</p> <p>This is story A: {StoryA}</p> <p>This is story B: {StoryB}</p> <p>Which of the stories A/B are analogous to the source story?</p> <p>Provide the final answer in tags: <ans>A </ans> or <ans>B </ans>.</p> <p>Only one answer is correct.</p>
Verbal Analogy	<p>The notation "A : B :: C : D" should be read as "A is to B like C is to D".</p> <p>In the following example:</p> <p>{A} : {B} :: {C} : ?</p> <p>Replace the question mark with the correct analogy: D: {D} or E: {D_prime}</p> <p>Provide the final answer in tags: <ans>A </ans> or <ans>B </ans>.</p> <p>Only one answer is correct.</p>

Table 2: **Basic prompt examples for story and verbal analogy tasks.** The content within the curly brackets is replaced with the actual content (story or word)

variants of the aforementioned prompts, resulting in twelve prompts in total. Table 3 shows an example of the basic prompt for each of the tasks. The remaining prompts and their variants can be found in the Appendix A and B. For every prompt type, we evaluate the majority voting with all three variants.⁶ Figure 1 presents our pipeline (the exact details of the steps taken for answer extraction are provided in the Appendix).

Second Stage Prompting In cases where the initial answer extraction from the model’s output was not parsable according to pre-defined rules, we implemented a second stage of prompting. This involved concatenating the original prompt with the initial output and appending the phrase "So the final answer is <ans>". This new prompt was then fed back into the model to generate a revised output. If the second stage parsing still did not yield the desired results, the response was treated as incorrect. The same extraction method was applied to the new output, and the majority vote was then performed over different prompt variants to obtain the final prediction.

5 Results and Analysis

This section presents the results of the experiments and provides an analysis of the findings. It includes comparisons between different models and prompt

types, and discusses any observed patterns or insights.

5.1 Story Analogy

Our experiments show (Figure 2; Table 3) that for all tested prompt types and all tested models, the observed accuracy is lower than the human baseline. This shows that the story analogy task remains difficult for the tested language models.

Another interesting observation is that the standard zero-shot CoT (addition of "Let’s think step by step") does not improve the accuracy when compared to the basic prompts. This result may be surprising, as CoT has been shown to improve results in many settings (Wei et al., 2022). This may suggest that unstructured CoT may lead to "overthinking" in models i.e., resulting in long reasoning that culminates in incorrect conclusion.

We also noted that, in some instances, the models did not actually perform reasoning before giving the final answer, even when using CoT, which can be another justification of the lack of improvement. This observation highlights inconsistencies in the models’ reasoning capabilities. For a more detailed description, please refer to the Appendix E.

However, we do observe that for Mistral and Starling, our CoT Structured prompt gives a significant improvement. This shows that for difficult tasks, like story analogy resolution, some human work in terms of preparing the reasoning plan is still required to get relatively good results. This

⁶We used it because we saw that it outperformed the cases when we used individual prompt types

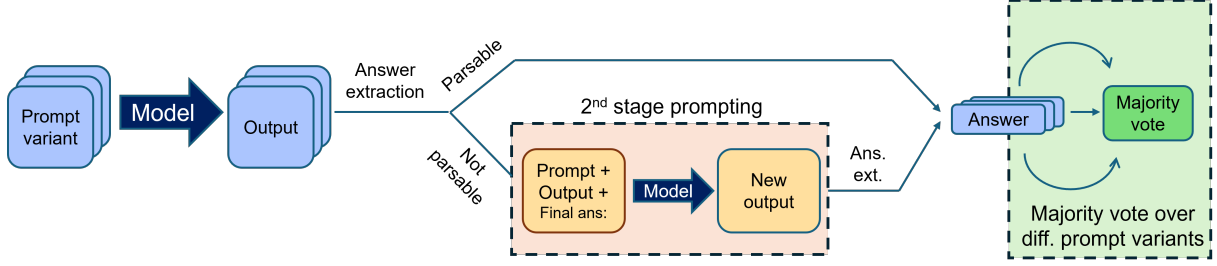


Figure 1: **Prompting pipeline.** Shows the iterative prompting process, including initial output generation, parsing, possible re-prompting, and final majority voting to achieve the prediction.

improvement, nevertheless, is not observed with LLaMa3. Upon closer inspection, we believe that structured CoT still resulted in an improvement, albeit one that is not reflected in the accuracy metric. The reasoning steps in the output of LLaMa3 when using structured CoT were often valid and captured the correct analogy; only the final answer was incorrectly concluded. We call such error cases "one-step away" errors (an example can be found in the Appendix D).

Overall, these results show that for 7-8B models, it may be insufficient to simply add "Let's think step by step" as the model might still arrive to the incorrect conclusion. Our findings underscore the importance of guiding the models on how to reason about the task, in order to improve performance.

5.2 Verbal Analogy

Our results demonstrate that both structured and unstructured CoT methods outperform basic prompt techniques in verbal analogy tasks. Notably, the standard CoT approach was sufficient to enhance performance beyond the baseline established by our basic prompt method. Furthermore, guiding

the model's thought process (structured CoT) before reaching a conclusion did not yield significant improvements over the standard zero-shot CoT approach.

We infer that CoT enhances performance in these tasks, likely due to the relative simplicity of verbal analogy tasks compared to the story analogy task. Additionally, all models surpassed the human baseline accuracy of 0.85, indicating their superior ability to understand semantic relations on a word level. Among the models tested, we lacked results for Starling across all prompt variants due to GPU unavailability. Therefore, our comparison focuses on LLaMA and Mistral, both of which performed comparably well, as illustrated in Table 3.

6 Conclusion

In this study, we investigated the effectiveness of zero-shot and structured Chain-of-Thought (CoT) prompting in medium-sized large language models (LLMs) for abstract analogy tasks. Our evaluation covered three models—LLaMA-3 8B, Starling 7B, and Mistral 7B—across story and verbal analogy tasks using a variety of prompt types.

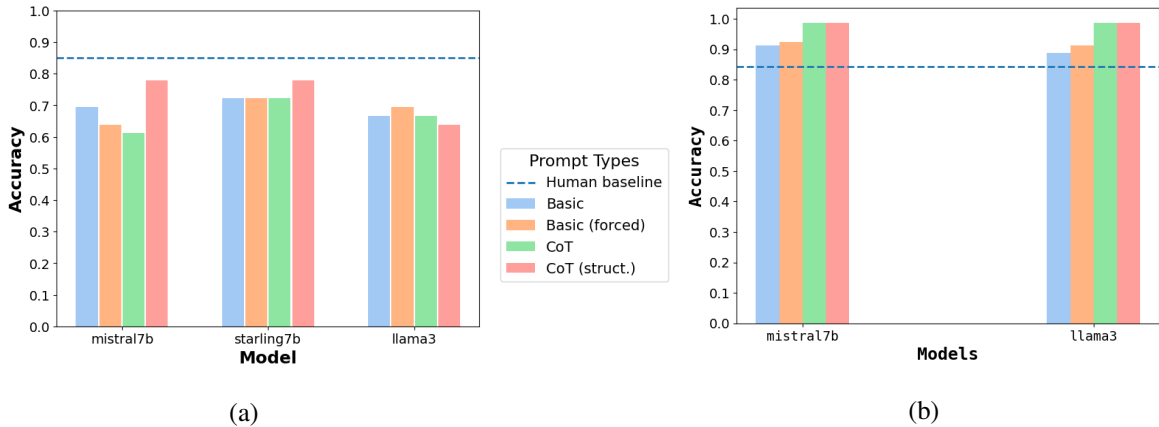


Figure 2: **Accuracy for different prompt types in (a) story analogy and (b) verbal analogy task** The dotted horizontal line indicates mean human performance. Starling couldn't be run due to GPU shortage for VA task.

Prompt type	Story Analogy			Verbal Analogy	
	Mistral	Starling	LLaMA	Mistral	LLaMA
Basic	0.69	0.72	0.67	0.91	0.88
Basic forced	0.64	0.72	0.69	0.92	0.91
CoT	0.61	0.72	0.66	0.98	0.98
CoT structured	0.78	0.78	0.64	0.98	0.98

Table 3: **Accuracy for different prompt types in verbal and story analogies for the different models** (Mistral and Starling are 7B in parameter, LLaMA is 8B in parameter).

Our findings indicate that standard zero-shot CoT prompting improves performance on verbal analogy tasks but does not enhance accuracy for story analogy tasks. This suggests that while CoT prompting can unlock reasoning capabilities in models, its effectiveness varies significantly depending on the complexity of the task. Interestingly, structured CoT prompting, which involves adding task-specific reasoning steps, showed a notable improvement in performance for story analogy tasks in two of the three models (Mistral and Starling). This highlights the importance of guiding models through structured reasoning processes for more complex tasks.

Overall, this research underscores the potential and limitations of CoT prompting in enhancing the abstract reasoning capabilities of medium-sized LLMs. While standard CoT techniques can significantly boost performance in some scenarios, careful design and task-specific adjustments through structured CoT are crucial for tackling more complex reasoning tasks effectively.

Limitations

In this paper, we observe that the methodology we propose shows the improved performance in English language, a language which has limited morphological complexity. We acknowledge that the models used were trained on large English corpus and hence the results align well with our model’s capabilities. It is important to acknowledge that this characteristic may limit the applicability of our method to more morphologically rich languages, suggesting a critical area for further research.

Moreover, we recognize additional constraints related to the scalability of our approach when processing longer texts. This challenge is compounded by the substantial computational resources required (for larger models), particularly in terms of GPU capacity. We believe that larger models may have

better ability to analyse the semantic relations in complex tasks like story analogy and hence might show better improved results.

Finally, we do note that the dataset on which we tested the model’s ability to capture verbal analogy might not have been the most challenging dataset available. The SCAN dataset (Czinczoll et al., 2022) provides examples that exhibit much more abstract relationships, quite challenging even for humans. We based our research on a set of tasks from (Webb et al., 2023), which is why we wanted to use the same datasets to ensure that our results are comparable. Testing the performance on a more challenging benchmark is highly desirable, but is left as a place for future work, given the scope of this project.

Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. The dataset used in the paper is publicly available. Additionally, authors understand that analogy evaluation relies on commonsense, and different models used are an attempt to capture the diverse perspectives.

Acknowledgements

We thank MSc Alina Leidinger for her inputs and guidance for throughout the project. We also thank Dr. Ekaterina Shutova for her lectures on CoT and LLM prompting which formed the basis of this report.

References

- Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. [Scientific and creative analogies in pretrained language models](#).
- Dedre Gentner, Mary Jo Rattermann, and Kenneth D Forbus. 1993. The roles of similarity in transfer:

Separating retrievability from inferential soundness. *Cognitive psychology*, 25(4):524–575.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. [Instruction tuned models are quick learners](#).

Douglas R. Hofstadter and Melanie Mitchell. 1994. *The Copycat project: A model of mental fluidity and analogy-making*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Hongjing Lu, Ying Nian Wu, and Keith J Holyoak. 2019. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181.

Mirac Suzgun, Nathan Scales, Nathanael Sch  rli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, et al. 2022. [Challenging big-bench tasks and whether chain-of-thought can solve them](#). *arXiv preprint*, arXiv:2209.12356.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. [Emergent analogical reasoning in large language models](#). *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaiif.

A Story Analogy Prompts

The following section lists all the prompts used for the task of story analogy.

A.1 Basic Prompt

A.1.1 Version 1

This is a source story:

{SourceStory}

This is story A:

{StoryA}

This is story B:

{StoryB}

Which of the stories A/B are analogous to the source story?

Provide the final answer in tags: <ans> A </ans> or <ans> B </ans>.

Only one answer is correct.

A.1.2 Version 2

Instruction:

You will be presented with one source story and two comparison stories.

Your task is to determine which of the comparison stories, Story A or Story B, is more analogous to the source story.

Only one answer is correct.

Source Story:

{SourceStory}

Story A:

{StoryA}

Story B:

{StoryB}

Answer:

Final answer should be in tags: <ans> A </ans> or <ans> B </ans>

A.1.3 Version 3

You are a social and humanities science expert, analyzing different stories and their similarity. You will be provided a reference story (Ref) and two other stories (A/B). You know that exactly one of the other two stories is analogous to the reference story. You are asked to select which of the other stories is analogous.

Ref: {SourceStory}

Story A: {StoryA}

Story B: {StoryB}

Provide the final answer in tags: <ans> A </ans>
or <ans> B </ans>.

A.2 Basic Prompt Forced

A.2.1 Version 1

This is a source story:

{SourceStory}

This is story A:

{StoryA}

This is story B:

{StoryB}

Which of the stories A/B are analogous to
the source story?

Return only the final answer in tags: <ans> A
</ans> or <ans> B </ans>.

Only one answer is correct.

A.2.2 Version 2

Instruction:

You will be presented with one source story and
two comparison stories. Your task is to determine
which of the comparison stories, Story A or Story
B, is more analogous to the source story.
Only one answer is correct.

Source Story:

{SourceStory}

Story A:

{StoryA}

Story B:

{StoryB}

Answer:

Final answer should be in tags: <ans> A </ans> or
<ans> B </ans>

Return only the final answer

A.2.3 Version 3

You are a social and humanities science expert,
analyzing different stories and their similarity. You
will be provided a reference story (Ref) and two
other stories (A/B). You know that exactly one of
the other two stories is analogous to the reference
story. You are asked to select which of the other
stories is analogous.

Ref: {SourceStory}

Story A: {StoryA}

Story B: {StoryB}

Provide the final answer in tags: <ans> A </ans>
or <ans> B </ans>.
Return only the final answer.

A.3 CoT

A.3.1 Version 1

This is a source story:

{SourceStory}

This is story A:

{StoryA}

This is story B:

{StoryB}

Which of the stories A/B are analogous to
the source story?

Provide the final answer in tags: <ans> A </ans>
or <ans> B </ans>.

Only one answer is correct.

Let's think step by step.

A.3.2 Version 2

Instruction:

You will be presented with one source story and

two comparison stories.

Your task is to determine which of the comparison stories, Story A or Story B, is more analogous to the source story.

Only one answer is correct.

Source Story:

{SourceStory}

Story A:

{StoryA}

Story B:

{StoryB}

Answer:

Final answer should be in tags: <ans> A </ans> or <ans> B </ans>

Let's think step by step.

A.3.3 Version 3

You are a social and humanities science expert, analyzing different stories and their similarity. You will be provided a reference story (Ref) and two other stories (A/B). You know that exactly one of the other two stories is analogous to the reference story. You are asked to select which of the other stories is analogous.

Ref: {SourceStory}

Story A: {StoryA}

Story B: {StoryB}

Provide the final answer in tags: <ans> A </ans> or <ans> B </ans>.

Let's think step by step.

A.4 CoT Structured

A.4.1 Version 1

This is a source story:

{SourceStory}

This is story A:

{StoryA}

This is story B:

{StoryB}

Which of the stories A/B are analogous to the source story?

1/ Briefly summarize and compare the themes of Story A and Story B with the Source Story.

2/ Evaluate how the relationships and development of characters in Story A and Story B align with those in the Source Story.

3/ Discuss the similarities and differences in the narrative structure between the stories.

4/ Based on the comparisons, determine which story (A or B) is more closely related to the Source Story.

5/ State the final answer. Provide final answer in tags: <ans> A </ans> or <ans> B </ans>

Only one answer is correct.

A.4.2 Version 2

Instruction:

You will be presented with one source story and two comparison stories.

Your task is to determine which of the comparison stories, Story A or Story B, is more analogous to the source story.

Only one answer is correct.

Source Story:

{SourceStory}

Story A:

{StoryA}

Story B:

{StoryB}

Answer:

Final answer should be in tags: <ans> A </ans> or <ans> B </ans>

Let's think step by step.

Guide: 1/ Briefly summarize and compare the themes of Story A and Story B with the Source Story.

2/ Evaluate how the relationships and development of characters in Story A and Story B align with those in the Source Story.

3/ Discuss the similarities and differences in the narrative structure between the stories.

4/ Based on the comparisons, determine which story (A or B) is more closely related to the Source Story.

5/ State the final answer. Provide final answer in tags: <ans> A </ans> or <ans> B </ans>

A.4.3 Version 3

You are a social and humanities science expert, analyzing different stories and their similarity. You will be provided a reference story (Ref) and two other stories (A/B). You know that exactly one of the other two stories is analogous to the reference story. You are asked to select which of the other stories is analogous.

Ref: {SourceStory}

Story A: {StoryA}

Story B: {StoryB}

Provide the final answer in tags: <ans> A </ans> or <ans> B </ans>.

1/ Briefly summarize and compare the themes of Story A and Story B with the Source Story.

2/ Evaluate how the relationships and development of characters in Story A and Story B align with those in the Source Story.

3/ Discuss the similarities and differences in the narrative structure between the stories.

4/ Based on the comparisons, determine which story (A or B) is more closely related to the Source Story.

5/ State the final answer. Provide final answer in tags: <ans> A </ans> or <ans> B </ans>

B Verbal Analogy Prompts

The following section lists all the prompts used for the task of verbal analogy.

B.1 Basic Prompt

B.1.1 Version 1

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Provide final answer in tags: <ans> D </ans> or <ans> E </ans>

Only one answer is correct.

B.1.2 Version 2

#Instruction:

You will be presented with one analogy and two potential completions.

Your task is to determine which of the completions, D or E, correctly completes the analogy.

Only one answer is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

Analogy:

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

#Answer

Final answer should be in tags: <ans> D </ans> or <ans> E </ans>

B.1.3 Version 3

You are a social and humanities science expert, analyzing different analogies and their correctness. You will be provided with one analogy and two potential completions (D/E).

You know that exactly one of the completions correctly completes the analogy. You are asked to select which of the completions is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Provide the final answer in tags: <ans> D </ans> or <ans> E </ans>.

B.2 Basic Prompt Forced

B.2.1 Version 1

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

Analogy: {A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Which of the completions D or E correctly completes the analogy?

Return only the final answer in tags: <ans> D </ans> or <ans> E </ans>.

Only one answer is correct.

B.2.2 Version 2

Instruction:

You will be presented with one analogy and two potential completions.

Your task is to determine which of the completions, D or E, correctly completes the analogy.

Only one answer is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

Analogy:

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Answer:

Final answer should be in tags: <ans> D </ans> or <ans> E </ans>

Return only the final answer

B.2.3 Version 3

You are a social and humanities science expert, analyzing different analogies and their correctness. You will be provided with one analogy and two potential completions (D or E).

You know that exactly one of the completions correctly completes the analogy.

You are asked to select which of the completions is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

Analogy:

$\{A\} : \{B\} :: \{C\} : ?$

Replace the question mark with the correct analogy:

D: $\{D\}$

or

E: $\{D_prime\}$

Provide the final answer in tags: `<ans> D </ans>` or `<ans> E </ans>`.

Return only the final answer.

B.3 CoT

B.3.1 Version 1

The notation " $A : B :: C : D$ " should be read as "A is to B like C is to D".

In the following example:

$\{A\} : \{B\} :: \{C\} : ?$

Replace the question mark with the correct analogy:

D: $\{D\}$

or

E: $\{D_prime\}$

Provide final answer in tags: `<ans> D </ans>` or `<ans> E </ans>`

Only one answer is correct.

Let's think step by step.

B.3.2 Version 2

#Instruction:

You will be presented with one analogy and two potential completions.

Your task is to determine which of the completions, D or E, correctly completes the analogy.

Only one answer is correct.

The notation " $A : B :: C : D$ " should be

read as "A is to B like C is to D".

In the following example:

Analogy:

$\{A\} : \{B\} :: \{C\} : ?$

Replace the question mark with the correct analogy:

D: $\{D\}$

or

E: $\{D_prime\}$

#Answer

Final answer should be in tags: `<ans> D </ans>` or `<ans> E </ans>` // // Let's think step by step.

B.3.3 Version 3

You are a social and humanities science expert, analyzing different analogies and their correctness. You will be provided with one analogy and two potential completions (D/E).

You know that exactly one of the completions correctly completes the analogy. You are asked to select which of the completions is correct.

The notation " $A : B :: C : D$ " should be read as "A is to B like C is to D".

$\{A\} : \{B\} :: \{C\} : ?$

Replace the question mark with the correct analogy:

D: $\{D\}$

or

E: $\{D_prime\}$ ---

Provide the final answer in tags: `<ans> D </ans>` or `<ans> E </ans>`.

Let's think step by step.

B.4 CoT Structured

B.4.1 Version 1

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Which of the completions D or E correctly completes the analogy?

1/ Briefly explain how completion D relates to the analogy.

2/ Briefly explain how completion E relates to the analogy.

3/ Compare the relationships presented in completions D and E with the analogy.

4/ Based on the comparisons, determine which completion (D or E) correctly completes the analogy.

5/ State the final answer. Provide final answer in tags: <ans> D </ans> or <ans> E </ans>

Only one answer is correct.

B.4.2 Version 2

#Instruction:

You will be presented with one analogy and two potential completions.

Your task is to determine which of the completions, D or E, correctly completes the analogy.

Only one answer is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

In the following example:

Analogy:

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

Guide:

1/ Briefly explain how completion D relates to the analogy.

2/ Briefly explain how completion E relates to the analogy.

3/ Compare the relationships presented in completions D and E with the analogy.

4/ Based on the comparisons, determine which completion (D or E) correctly completes the analogy.

5/ State the final answer. Provide final answer in tags: <ans> D </ans> or <ans> E </ans>

Only one answer is correct.

B.4.3 Version 3

You are a social and humanities science expert, analyzing different analogies and their correctness. You will be provided with one analogy and two potential completions (D/E).

You know that exactly one of the completions correctly completes the analogy. You are asked to select which of the completions is correct.

The notation "A : B :: C : D" should be read as "A is to B like C is to D".

{A} : {B} :: {C} : ?

Replace the question mark with the correct analogy:

D: {D}

or

E: {D_prime}

1/ Briefly explain how completion D relates to the analogy.

2/ Briefly explain how completion E relates to the analogy.

3/ Compare the relationships presented in completions D and E with the analogy.

4/ Based on the comparisons, determine which completion (D or E) correctly completes the analogy.

5/ State the final answer. Provide final answer in tags: <ans> D </ans> or <ans> E </ans>

C Story Analogy Full Example

Table 4 shows a single sample from the story analogy dataset for illustration of the problem.

D One-Step Away Errors

This section shows an example of a one-step away error i.e., the case when the model’s reasoning was correct all the way up to the final conclusion. This was often observed with structured Chain-of-Thought prompting. While the final answer is still incorrect, this is much closer to the correct solution compared to the case of the basic prompt (not forced). Previous work (Kojima et al., 2022) show that similar errors types were often resolved by increasing the model size. Table 5 shows an example story in which we observed a one-step away error in the generated output and Table 6 shows the output of LLaMa3.

We can observe that the model found the correct analogy i.e., “a clear setup with a conflict and resolution” and “linear progression, with the accident serving as a turning point”. However, it still fails to map this analogy to the correct story in the final step.

E Answer Position Analysis

Through analysing the results, we noted substantial variability in model performance across different answer positions (beginning, middle, end), reflecting a potential correlation between the models’ training paradigms and their information processing capabilities. Particularly, the data shows a pronounced impact of answer positioning on model accuracy, with a noticeable decrement in the performance at middle positions (see Figure 3). However, due to time constraints, a comprehensive exploration of these aspects was not included in the current scope of the analysis, and should be pursued in future research.

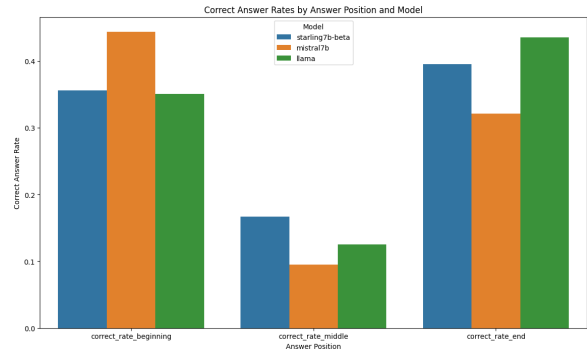


Figure 3: Accuracy on Story Analogies Task by Model and Position

Story Analogy	
Source Story	A dog named Leonard was trotting along a dock holding a steak in his teeth. However, as he trotted he dropped it inadvertently and it disappeared into the lake. Thus, he was left without his dinner. Then Leonard saw another dog with a steak. Because he had lost his own steak he tried to take it. Unluckily, the other dog was much bigger than he was. Leonard ended up losing a tooth in the scuffle.
Story A	Charlie was walking through a hotel with an expensive pocket watch in his hand. As he walked he accidently dropped it and it shattered on the floor. Consequently, he was left without a watch. Then Charlie saw another man with an expensive watch. Since he had broken his own watch he reached out to take it. Unfortunately, the man was a policeman.
Story B	Charlie was walking through a hotel with an expensive pocket watch in his hand. Then he noticed another man with an expensive watch. Not realizing it was just his reflection, and since he was also quite greedy, Charlie reached out to take it. Unfortunately, as he opened his hand he dropped his own watch and both watches shattered on the floor. Consequently he was left with no watch at all.

Table 4: **Example sample from the dataset for story analogy.**

One-Step Away Error	
Source Story	<p>Mr. Newton was the manager of a company that made razors. One year, an inventor in his company perfected a metal that would stay sharper than any previous razor. However, not long after that, Mr. Newton was using one of these new razors and he cut his neck rather severely. As a result of this incident the inventor recommended a safety feature that would protect people from injuring themselves. After what had happened, Mr. Newton considered the safety feature absolutely necessary. The incident had alarmed him so much that he began to produce the safety razor right away. Luckily, his neck did not get infected so everything was alright.</p>
Story A	<p>Ms. Boyce was in charge of installing sun lamps sold by Smith Tan Company. In 1980, a supplier for the company perfected a filament that could produce more intense heat than any previous filament. Because it was so hot, the supplier recommended a warning label that would protect people from hurting themselves with the new lamp. But Mrs. Boyce considered the labels unnecessary. It would be too much trouble to make them. But later, when she was using one of these new heat lamps herself, she was seriously burned. After this accident she was so frightened that she had the warning labels printed immediately.</p>
Story B	<p>Ms. Boyce was in charge of installing heat lamps sold by the Smith Tan Company. In 1980, a supplier for the company perfected a filament that could produce more intense heat than any previous filament. But later, when Ms. Boyce was using one of these new heat lamps herself she was seriously burned. Because of this accident the supplier recommended a warning label that would protect people from hurting themselves with the new lamp. After this, Ms. Boyce considered the labels absolutely necessary. The accident frightened her so much that she had the labels printed immediately.</p>

Table 5: **Particular data sample from the dataset where the model’s generation exhibited a one-step away error.**

One-Step Away Error

Summary and Comparison of Themes

The Source Story, Story A, and Story B all revolve around a character (Mr. Newton/Ms. Boyce) who is involved in a company that produces a product (razors/heat lamps) and has an accident while using the product. The themes of the stories are similar, focusing on the importance of safety features and the consequences of not prioritizing them.

Evaluation of Character Relationships and Development

In the Source Story, Mr. Newton is the manager of the company and is directly involved in the development of the new razor. He is also the one who has the accident and is subsequently convinced of the need for a safety feature. In Story A, Mrs. Boyce is also directly involved in the installation of the new heat lamps and has the accident. However, in Story B, the supplier recommends the warning label after the accident, which is a less direct involvement. The character development in all three stories is similar, with the main character being alarmed and convinced of the need for a safety feature after the accident.

Model output**Comparison of Narrative Structure**

The narrative structure of all three stories is similar, with a clear setup, conflict, and resolution. The stories follow a linear progression, with the accident serving as a turning point that leads to the main character's realization of the importance of safety features.

Comparison and Conclusion

Based on the comparisons, Story A is more analogous to the Source Story. Both stories feature a character who is directly involved in the development of the product, has an accident, and is subsequently convinced of the need for a safety feature. The narrative structure and character development are also similar. Story B, while sharing some similarities, has a less direct involvement of the main character in the development of the product and the recommendation of the safety feature.

Final Answer

<ans> A </ans>

Table 6: One-step away error.