Can Stories Help LLMs Reason? Curating Information Space Through Narrative

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Narrative is widely recognized as a powerful tool for structuring information and facilitating comprehension of complex ideas in various domains, such as science communication. This paper investigates whether incorporating narrative elements can assist Large Language Models (LLMs) in solving complex tasks more effectively. We propose a novel approach, **Story of Thought (SoT)**, integrating narrative structures into prompting techniques for problem-solving tasks. This approach involves constructing narratives around problem statements and creating a framework to identify and organize relevant information. We hypothesize that this narrative-based information curation process enhances problem comprehension by contextualizing critical information and highlighting causal relationships within the problem space. Our experimental results show that the SoT approach consistently surpasses Chain of Thought (CoT) and Analogical Reasoning in GPQA tasks, achieving higher accuracy and better solutions in physics, chemistry, and biology problem-solving tasks with all tested OpenAI, Meta, and Mistral LLMs.

1 Introduction

Humans have an exceptional ability to understand and reason through narratives. A narrative-driven approach can enhance the comprehension and retention of complex subjects compared to simple fact listing Fisher [2021], Abbott [2020], Gottschall [2012]. For example, storytelling effectively structures information in science communication Dahlstrom [2014], Norris et al. [2005], Martinez-Conde and Macknik [2017], education Engel et al. [2018], Negrete and Lartigue [2004], and health communication Dudley et al. [2023], revealing relationships and contextual nuances Zak [2015]. As shown in Figure 1, the *factual approach* presents information in a concise manner akin to a reference source, whereas the *narrative approach* conveys information through storytelling to contextualize facts within a broader setting.

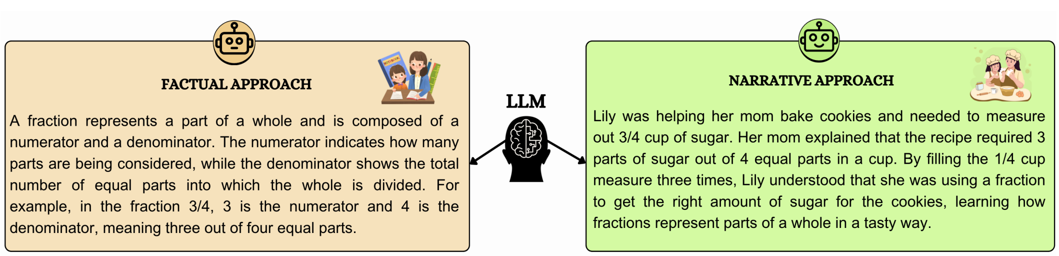

Figure 1: Contrasting approaches to information delivery, illustrated on explaining the concept of fractions: Factual vs. Narrative.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

To date, large language models (LLMs) struggle with complex problem-solving tasks that require the ability to integrate, structure, and apply relevant information effectively Qiao et al. [2023], Wang et al. [2023]. Prompting techniques based on breaking tasks into smaller subtasks, such as Chain-of-Thought (CoT) Wei et al. [2022] and its more recent adaptations Xia et al. [2024], have led to considerable improvements in problem-solving benchmarks. The strategies of constructing natural language rationales Ling et al. [2017], in the CoT context also called reasoning processes, play a vital role in LLM prompting Ye and Durrett [2024], Min et al. [2022], Wang et al. [2022], Li et al. [2023].

Inspired by the effectiveness of narrative in *(i)* identifying and explaining important concepts and *(ii)* organizing the information flow coherently, we explore integrating narrative elements into prompt-driven reasoning. The main research questions addressed in this work are:

RQ 1: Can LLMs generate coherent and relevant narratives around problem statements to facilitate problem comprehension and reasoning?

RQ 2: Can incorporating narrative elements into LLM prompting techniques improve their performance on complex problem-solving tasks?

We make the following contributions to the RQs:
(i) We introduce a novel method, called **Story of Thought (SoT)**, that aids LLMs to identify and arrange relevant information for solving complex tasks by incorporating narrative structures into the prompting process. *(ii)* We evaluate the effectiveness of SoT on diverse, complex tasks, including physics, chemistry, and biology problem-solving in GPQA, showing superior performance to existing task-decomposition-based prompting techniques, such as zero-shot and few-shot chain of thought and analogical reasoning. *(iii)* We analyze the impact of individual narrative techniques used on the generated narrative-based explanation to investigate why they improve LLMs reasoning capabilities.

2 Related Work

Narrative *Narrative*, as a noun, refers to a story or a description of a series of events [1]. In other words, it is a particular way of explaining or understanding events and plays a crucial role in human communication and cognition Hineline [2018]. The terms "story" and "narrative" are often used interchangeably. However, there is a subtle difference between them. A "story" is a narrative's content or substance, while a "narrative" is the structure or way the story is presented Abbott [2020]. Narrative plays a crucial role in various aspects of human communication and cognition. Bruner [1991] argues that narrative is a fundamental mode of human thought, allowing individuals to organize and make sense of their experiences. However, there are also potential disadvantages to using narrative. One concern is that narratives can oversimplify complex issues or events, leading to a reductionist understanding Dahlstrom and Ho [2012]. Furthermore, an over-reliance on narrative structures may limit the exploration of alternative viewpoints or non-linear forms of information presentation Negrete and Lartigue [2004]. It is essential to balance the benefits of narrative and the need for a nuanced understanding of the problem space Avraamidou and Osborne [2009].

Narrative and Human Cognition Research into *narrative transportation* examines how individuals become cognitively and emotionally absorbed in stories. This immersive experience enhances emotional responses and alters attitudes and beliefs by aligning the listener's brain with the storyteller's Oschatz and Marker [2020], Bilandzic et al. [2020]. Neuroimaging techniques such as functional Magnetic Resonance Imaging (fMRI), Positron Emission Tomography (PET), and Magnetoencephalography (MEG) show differences in identified active brain regions involved in narrative comprehension compared to factual processing Sanford and Emmott [2012], Armstrong [2020], Aboud et al. [2019], Coopmans and Cohn [2022]. Furthermore, presenting information in narratives can enhance learning and memory, as well as promote engagement and motivation Willingham [2004], Rowe et al. [2010], Chen et al. [2023], which led to the development of narrative-based educational strategies Bower and Clark [1969], Mawasi et al. [2020], Norris et al. [2005].

Role of Narrative in Solving Tasks In problem-solving, narratives can serve as a framework for organizing and presenting information relevant to the task Jonassen and Hernandez-Serrano [2002],

[1] https://dictionary.cambridge.org/

Andrews et al. [2009]. Structuring the problem space as a narrative makes it easier to identify key elements, such as characters, goals, obstacles, and potential solutions San Roque et al. [2012]. Narratives can also help to break down complex problems into sub-problems, providing a step-by-step approach to problem-solving Szurmak and Thuna [2013]. For example, progressive disclosure, analogy, and analogical reasoning are powerful narrative techniques that facilitate problem-solving Salvucci and Anderson [2001], Gick and Holyoak [1980]. These techniques involve presenting information gradually in sub-problems, drawing comparisons and similarities between two seemingly disparate concepts, and using these similarities to generate insights or solutions Norris et al. [2005], Holyoak and Thagard [1989].

Narrative and LLM Prompting The intuitive approach to improving reasoning with LLMs is prompt engineering (see recent survey in Qiao et al. [2023]. Starting from CoT prompting Wei et al. [2022], these techniques leverage LLMs' strong in-context learning ability, adding intermediate steps to generate a reasoning process before answering. While analogical reasoning can be a technique used in narrative generation, to our knowledge the narrative technique has never been fully explored as a coherent set of interconnected didactic approaches to improve the reasoning abilities of LLMs for problem-solving.

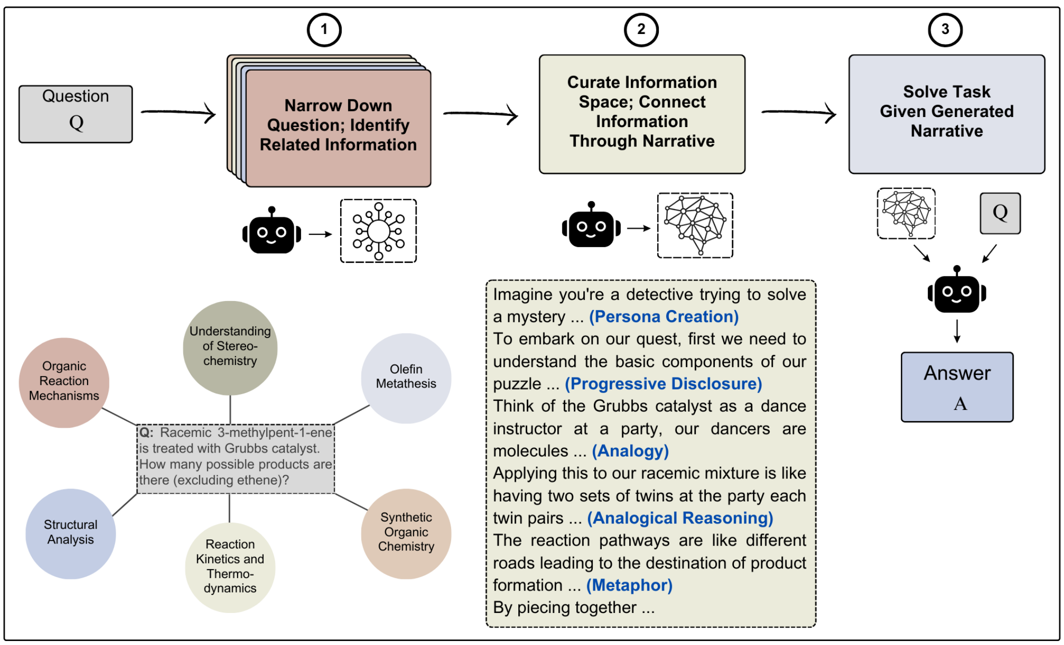

Figure 2: A high-level overview of **Story of Thought** (SoT), consisting of three steps (top): ① Question Clarification, ② Narrative Generation, ③ Solving Task and an actual example of LLM output (bottom) in each step for the GPQA task. The prompt designed for step 2 incorporates the narrative techniques (highlighted in blue) such as *analogical reasoning*, which identifies similarities between the target concept (information being conveyed) and a more familiar concept (*analogy*) and *progressive disclosure* which reveals information gradually throughout the narrative, rather than presenting it all at once. See Appendix B for the complete prompt for each step. See Appendix A for a complete example.

3 Methodology

We introduce **Story of Thought** (SoT), a novel prompt-driven reasoning approach that generates narrative-based clarification to guide LLMs' reasoning process. Inspired by the narrative format, the SoT approach leverages the cognitive benefits of storytelling, such as contextual understanding and relational reasoning, that can help LLMs identify and maintain the information structure.

Figure 2 gives an overview of SoT. It involves three steps using narrative techniques: *(i)* **Question Clarification** (i.e., acting as an explorer to dissect and clarify complex questions (Section 3.1)); *(ii)* **Narrative Generation** (i.e., generating detailed narratives from the clarified question components using different narrative techniques (Section 3.2)); and *(iii)* **Solving Task** (i.e., leveraging narratives to prompt the LLMs to solve the tasks (Section 3.3)).

3.1 Step 1: Question Clarification

In the first step, we use the LLM's ability to explore and clarify the problem. Starting with a specialized prompt, the LLM breaks down the question into its core components, identifying relevant subtopics and areas. This detailed analysis is crucial for generating a coherent narrative that thoroughly addresses the question. The prompt is shown in Appendix B.1.

3.2 Step 2: Narrative Generation

The second step involves generating detailed narratives based on the breakdown and clarification performed in Step 1 (i.e., Question Clarification described in the previous section). These narratives provide a structured context for the questions to enhance the LLM's understanding, reasoning, and problem-solving abilities. In synthesizing the literature discussed in Section 2, we integrate the following narrative techniques into our prompt below and task LLM to generate a narrative, based on the information identified in Step 1 (See Appendix B.2 for designed prompt):

1. **Progressive Disclosure**: Reveals information gradually, guiding the LLM step-by-step through the problem-solving process.
2. **Branching**: Explores different paths or approaches to understanding the problem by providing multiple perspectives.
3. **Analogy**: Uses comparisons to familiar concepts or situations to make abstract components more understandable.
4. **Analogical Reasoning**: Facilitates understanding by reasoning through similarities between the problem and known situations.
5. **Metaphor**: Simplifies complex ideas through metaphorical representation.

3.3 Step 3: Solving Task

In the final step, the LLM uses the narrative generated in Step 2 to solve the original QA task. The structured and contextual understanding provided by the narrative supports LLM in accessing relevant aspects of the task. (The prompt is shown in Appendix B.3)

4 Experimental Setup

To comprehensively evaluate the effectiveness of our proposed approach, we conduct experiments across a diverse set of tasks and models, employing various prompting techniques for comparison.

4.1 Evaluation Tasks

We focus our evaluation on reasoning-intensive tasks spanning multiple domains, including physics, biology, and chemistry problem-solving. In particular, we utilize the **GPQA** (Diamond set), a Graduate-level Problem-solving QA dataset Rein et al. [2023], which comprises expert-crafted multiple-choice questions. These tasks are diverse and extremely challenging, requiring in-depth reasoning and domain knowledge, making them well-suited for assessing our approach's ability to understand complex tasks and contextualize salient information within the problem space.

4.2 Evaluated Large Language Models

To evaluate the performance of our approach across a wide range of Large Language Models, we experiment with the following LLM families:

1. Meta: Llama 3 8B & Llama 3 70B **2. Mistral**: Mistral 7B & Mixtral 8x7B

3. OpenAI: GPT-3.5-turbo & GPT-4 **4. Microsoft**: Phi 3 Medium & Phi 3 Mini

Table 1: Performance (QA accuracy) of LLMs across prompting methods on GPQA (Diamond set).

Prompting Method	**Meta**		**Mistral**		**OpenAI**		**Microsoft**	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mixtral 8x7B	ChatGPT 3.5	GPT 4	Phi-3 Mini	Phi-3 Medium
Zero-shot	34.2	39.5	35.8	36.36	30.6	34.7	**28.79**	42.42
Zero-shot CoT	40.91	41.92	31.82	35.35	28.1	35.7	24.75	39.39
Analogical Reasoning (3-shot)	40.91	47.47	37.9	26.26	28.1	41.41	16.67	**48.48**
Ours: Knowledge Identification	40.4	48.99	35.35	37.77	27.77	40.90	20.71	37.88
Ours: Story of Thought (SoT)	**43.43**	**51.01**	**38.4**	**38.89**	**30.8**	**48.98**	22.73	36.36

These models were selected to cover a wide spectrum of capabilities and sizes, enabling a comprehensive evaluation of their strengths and limitations. By including models from multiple leading AI research organizations, we aim to provide a balanced comparison.

All experiments, except for those involving OpenAI models, were conducted on local machines equipped with GPUs. The models were run locally on a GPU setup without quantization using the Hugging Face Transformer library[2]. For OpenAI's GPT-3.5-turbo and GPT-4 models, we use the OpenAI API to generate outputs. Across all models, we use a temperature of 1.0 and a maximum number of tokens of 8,000 and report the accuracy.

4.3 Prompting Methods Benchmarked

We compared our proposed approach against several prompting techniques, including:

Zero-shot Prompting: This method, similar to our approach (SoT), does not rely on labeled examples. Instead, LLMs are prompted to solve tasks based solely on their pre-trained knowledge without any context provided. This approach serves as a baseline, demonstrating the LLMs' ability to solve problems without explicit guidance.

Zero-shot CoT Wei et al. [2022]: This technique extends the zero-shot prompting approach by encouraging the LLM to explicitly reason through the steps required to arrive at an answer. By prompting the model to generate a chain of thought, this method aims to improve the model's ability to solve complex problems by breaking them down into smaller, more manageable steps.

Analogical Reasoning Yasunaga et al. [2023]: This approach leverages analogies to help the model draw parallels between known concepts and the task at hand. By providing analogical examples, the model is guided to understand and apply similar reasoning patterns to new problems. In our experiment, we allow the LLMs to self-generate three exemplars for each question (akin to the prompt described in their paper). This enables them to identify relevant examples and adapt their reasoning accordingly.

Ours: Knowledge Identification: To measure the effectiveness of our proposed approach, namely utilizing narrative in solving tasks, we prompt LLMs to solve the task based solely on the generated conceptual knowledge from Step 1 (described in Section 3.1). This allows us to compare the model capability in solving tasks using only the identified relevant knowledge versus leveraging this knowledge to structure a coherent narrative.

Ours: Story of Thought (SoT): This approach represents the core of our proposed method, where we leverage the generated narratives from Step 2 (described in Section 3.2) to solve the given tasks.

5 Results

5.1 Benchmark Performance Results

The main results of our experiments on the GPQA task are presented in Table 1. We evaluate the performance of various prompting methods across eight different LLMs from four major AI companies: Meta, Mistral, OpenAI, and Microsoft. Our proposed prompt-driven reasoning approach (SoT), consistently outperformed the baseline approaches, including zero-shot prompting, zero-shot Chain-of-Thought (CoT) prompting, and analogical reasoning, for six out of the eight LLMs tested. This finding highlights the potential of leveraging narrative structures to improve the ability of LLMs to understand and reason about the given information in various intensive-reasoning tasks across a range of models. In particular, the open-source Llama 3 70B model records the highest accuracy

[2] `https://huggingface.co/docs/transformers`

using the SoT method, achieving a score of 51.01%. This is the highest accuracy observed among all models and methods tested in the study, and, at the time of writing also a state-of-the-art result compared to public leaderboards (including e.g., Claude 3 and Gemini 1.5[3]). Furthermore, the GPT-4 model shows the most notable improvement in accuracy when the SoT method is employed, compared to its zero-shot baseline. Specifically, the accuracy for GPT-4 increased from 34.7% under zero-shot conditions to 48.98% with SoT (i.e., an absolute increase of 14.28%, or a relative increase of 41% respectively).

Interestingly, all reasoning strategies lead to an accuracy drop for the comparably smaller Phi-3 Mini model, and all CoT strategies except Analogical Reasoning also lead to the accuracy drop of the Phi-3 Medium model compared to its zero-shot baseline. We hypothesize that this is due to the low quality of the generated explanations (whether CoT steps or SoT narrative), as further indicated in the following subsection.

5.2 Role of the Narrative Quality/Choice

We further investigate the role of the choice of *narrator* model (i.e., the model that generates narratives) for problem-solving tasks. In the following experiments, we apply the narratives generated by other large and small open-source LLMs to the Phi-3 Mini and Phi-3 Medium models. The results of these experiments are presented in Table 2.

We observe that the narratives generated by the Llama 3 8B, Llama 3 70B, and Mistral 7B models consistently improve the accuracy of both Microsoft models compared to the baseline (i.e., when both models use their own generated narratives in Step 2 to solve the tasks, shown in Table 1). The absolute improvements range from 1.0% to 2.8%, with the Llama 3 70B model generating the most effective narratives. A slight decrease in accuracy is observed with the mixture-of-experts Mixtral 8x7B narratives for the Phi-3 Medium model, highlighting the need for careful selection and evaluation of narrator models to ensure compatibility and optimal performance.

Table 2: Applying generated narratives by open-source models to Microsoft models to solve the tasks.

Narrative Generator	**Solver Models**	
	Phi-3 Mini	Phi-3 Medium
Llama 3 8B	23.74 (+1.01↑)	37.88 (+1.28↑)
Llama 3 70B	25.25 (+2.52↑)	**39.39** (+2.79↑)
Mistral 7B	24.24 (+1.51↑)	38.38 (+1.78↑)
Mixtral 8x7B	24.74 (+2.01↑)	35.86 (-0.74↓)

5.3 Impact of Narrative Elements

To measure the impact of each of the five individual narrative techniques, we jointly prompted on the performance of open-source Meta models, we ablate the designed prompt in Step 2 (of Section 3.2) to apply each of the techniques separately. The results in Table 3 indicate that employing any single narrative technique at a time is notably less effective at boosting QA accuracy than utilizing a combination of these simultaneously.

Table 3: Comparing accuracy when using a single narrative technique. The values in parentheses represent the decrease in accuracy percentage points compared to a combination of multiple narrative techniques simultaneously (shown in Table 1).

Narrative Technique	**Meta**	
	Llama 3 8B	Llama 3 70B
Progressive Disclosure	34.85 (-8.58↓)	44.95 (-6.06↓)
Branching	34.34 (**-9.09**↓)	44.95 (-6.06↓)
Analogy	39.39 (-4.04↓)	46.46 (-4.55↓)
Analogical Reasoning	40.4 (-3.03↓)	45.45 (-5.56↓)
Metaphor	41.41 (-2.02↓)	44.44 (-6.57↓)
All	43.43	51.01

[3] `https://klu.ai/glossary/gpqa-eval`

For both models (Llama 3 8B and 7B), the decrease in accuracy is comparably smaller (-3.0% to -5.6%) when using only the analogical components of the narrative (*Analogy* and *Analogical Reasoning*) than when using only the structural instructions (*Progressive Disclosure* or *Branching*) which leads to larger (-6.0% to -9.1%) accuracy loss.

However, reasoning alone does not perform on par with the full narrative generation listing all the techniques. Prompting for *Metaphor* usage only leads to a larger accuracy loss in the 70B model (-6.6%) compared to the smaller one ((-2.0%). This makes us wonder to which extent the narrative techniques are correlated, and to which extent the model can "understand" what it is prompted for, which we attempt to analyze in the following subsections.

5.4 Analyzing Generated Narratives

To gain deeper insights into the generated narratives, we designed a prompt (shown below) that utilizes our best-performing model (LLama 3 70B) to annotate the number of occurrences of each narrative technique for each generated narrative by all models used in our experiments. The intuition behind this experiment is that we can better interpret how the model executed the narrative technique prompt, by asking it to label if and where the mentioned techniques are used in the text generated. Less frequently labeled techniques might be the ones where LLM doesn't have a clear understanding of what it is asked to do. A proportion of the techniques and their correlation can provide us with a better picture of LLM's interpretation of the instruction as well.

Table 4: Comparing Generated Narratives - Total Number of Occurrences for each Narrative Techniques (Evaluator: Llama 3 70B)

Narrative Technique	Meta		Mistral		OpenAI		Microsoft	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mixtral 8x7B	ChatGPT 3.5	GPT 4	Phi-3 Mini	Phi-3 Medium
Progressive Disclosure	427	597	191	191	744	570	367	368
Branching	30	56	51	20	72	168	34	61
Analogy	418	425	117	161	498	595	569	499
Analogical Reasoning	205	191	78	108	213	336	276	206
Metaphor	249	316	103	137	811	428	418	291
$\sum$	1329	1585	540	617	2338	2097	1664	1425

We aim to uncover patterns and variations in the use of narrative techniques across different LLMs. Table 4 indicates a comparison of the total number of occurrences for each narrative technique across various LLMs.

Variability in Utilization of Narrative Techniques Across Models: In our designed prompt in Step 2 (i.e., Narrative Generation, described in Section 3.2), we task LLMs to generate narrative using all of the 5 narrative techniques. However, as Table 4 indicates, not all techniques were employed equally. The result reveals that while some techniques like *Analogy* and *Progressive Disclosure* were consistently utilized, others such as *Branching* were applied less frequently.

We observe a trend across all LLM families where models with larger capacities, such as Llama 3 70B and GPT-4, consistently show higher occurrences of narrative techniques compared to their smaller counterparts. Furthermore, OpenAI's models (ChatGPT 3.5 & GPT-4) demonstrate the highest total occurrences of narrative techniques, with 2,338 and 2,097, respectively with a notable emphasis on *Metaphors* and *Analogies*.

Correlation Among Narrative Techniques: To further investigate the dynamics of narrative techniques, we compute correlations between the frequencies of narrative techniques across solved and unsolved tasks, as shown in Figure 3. This analysis aims to uncover if the models consistently use certain narrative techniques together or vary significantly. Our initial results indicate diverse correlation patterns, suggesting that the effectiveness of narrative techniques in solving tasks across various LLMs needs to be further analyzed.

6 Limitations

Contribution limitations. The occurrences of narrative techniques do not necessarily imply the quality or effectiveness of the generated narratives; rather, they provide insights into the models' tendencies and preferences in employing these techniques. Therefore, answering the question of

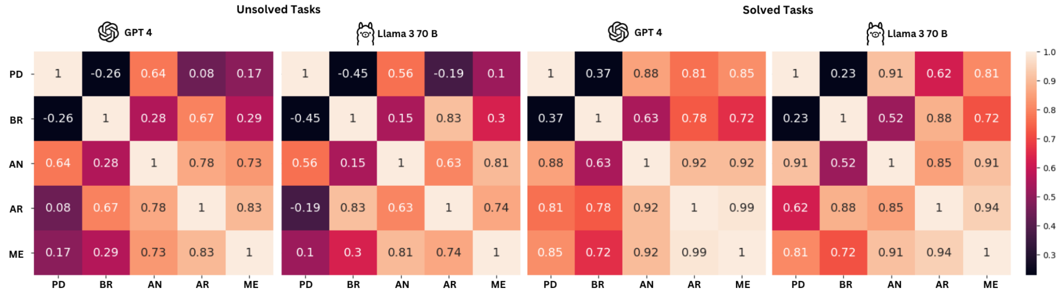

Figure 3: Correlation coefficients among all narrative techniques (**PD** = Progressive Disclosure, **BR** = Branching, **AN** = Analogy, **AR** = Analogical Reasoning, **ME** = Metaphor) used in the SoT approach for GPT-4 and Llama 3 70 B in solved and unsolved tasks.

Table 5: Performance of various LLMs across different prompting methods on GPQA (Diamond set). Correct answers are presented in the second option. Values in parentheses indicate the change in accuracy compared to the original setting in Table 1 where the correct answer was in the first option.

Prompting Method	**Meta**		**Mistral**		**Microsoft**	
	Llama 3 8B	Llama 3 70B	Mistral 7B	Mixtral 8x7B	Phi-3 Mini	Phi-3 Medium
Zero-shot	30.81 (-3.39↓)	31.31 (-8.19↓)	19.7 (-16.1↓)	18.18 (-18.18↓)	29.8 (+1.01↑)	21.72 (-20.7↓)
Zero-shot CoT	27.27 (-13.64↓)	33.33 (-8.59↓)	**22.73** (-9.09↓)	17.17 (-18.18↓)	32.32 (+7.57↑)	21.21 (-18.18↓)
Analogical Reasoning	27.78 (-13.13↓)	40.91 (-6.56↓)	10.61 (-27.29↓)	19.19 (-7.07↓)	**35.86** (+19.19↑)	16.67 (-31.81↓)
Ours: Knowledge Identification	32.32 (-8.08↓)	42.4 (-6.59↓)	16.67 (-18.68↓)	14.65 (-23.12↓)	28.28 (+7.57↑)	23.26 (-14.62↓)
Ours: Story of Thought (SoT)	**34.85** (-8.58↓)	**45.4** (-5.61↓)	20.2 (-18.2↓)	**20.2** (-18.69↓)	27.7 (+4.97↑)	**25.75** (-10.85↓)

why narrative is helping LLMs is more complex and needs to be further investigated by looking into different research areas such as cognitive and communication theories.

Method limitations. This method might not be efficient for tasks such as the MMLU benchmark, where the answer to the question depends on the provided options, because part of the information necessary to determine the correct answer is contained within the options themselves. To address this, we may include the options' information as part of the question, thereby ensuring that all relevant information is available for the method to process and derive the correct answer.

Dataset limitations. So far, we used only GPQA tasks as the most challenging set of problem-solving benchmarks we were aware of. Other comparable benchmarks, such as MGSM, are much closer to human or superhuman accuracy already without reasoning prompts and will be explored in future work.

Analysis limitations. We used Llama 70 B to respectively analyze the narratives. The intuition behind this experiment is that we can better interpret how the model executed the narrative technique prompt, by asking it to label if and where the mentioned techniques are used in the text generated. An alternative would be a thorough human assessment and further analysis of the impact on downstream performance, both of which we pursue in ongoing follow-up experiments. (We also previously prompted the LLMs in Step 2 to explain each of these five narrative techniques to make sure the concepts are understood before generating the narrative.)

LLM Robustness Limitations (Position of Correct Option). In the original GPQA dataset used for our experiments, the correct answers are always presented as the first option among the multiple choices. However, To further evaluate the robustness of the LLMs, we conduct an additional experiment where the correct answers are placed in the second option instead. Table 5 presents the results of these experiments, comparing the performance of various prompting methods across six different open-source LLMs. We observe that most LLMs experience a significant drop in accuracy when the correct answer is moved to the second option. However, despite the overall decrease in accuracy, our proposed approach, Story of Thought (SoT), consistently outperforms the baseline methods for most LLMs. The SoT method achieves the highest accuracy for the Meta Llama 3 8B, Meta Llama 3 70B, Mistral 8x7B, and Microsoft Phi-3 Medium models, demonstrating its effectiveness in enhancing the robustness of LLMs to changes in the problem structure.

7 Conclusions

Inspired by findings from human cognitive processes explored in didactics research, in this work, we propose to use narrative techniques in LLM prompting.We present strong evidence on public benchmark datasets that narrative techniques have the potential to notably enhance the reasoning abilities of LLMs in complex problem-solving tasks. By incorporating narrative structures, which mimic human cognitive processes of organizing and interpreting information, LLMs can achieve higher levels of performance and provide more contextually enriched responses.

References

H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2020.

Katherine S Aboud, Stephen K Bailey, Stephanie N Del Tufo, Laura A Barquero, and Laurie E Cutting. Fairy tales versus facts: Genre matters to the developing brain. *Cerebral Cortex*, 29(11): 4877–4888, 2019.

Dee H Andrews, Thomas D Hull, and Jennifer A Donahue. Storytelling as an instructional method: Descriptions and research questions. 2009.

Paul B Armstrong. *Stories and the brain: The neuroscience of narrative*. Johns Hopkins University Press, 2020.

Lucy Avraamidou and Jonathan Osborne. The role of narrative in communicating science. *International Journal of Science Education*, 31(12):1683–1707, 2009.

Helena Bilandzic, Susanne Kinnebrock, and Magdalena Klingler. The emotional effects of science narratives: a theoretical framework. *Media and Communication*, 8(1):151–163, 2020.

Gordon H Bower and Michal C Clark. Narrative stories as mediators for serial learning. *Psychonomic science*, 14(4):181–182, 1969.

Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.

Althea Y Chen, Chun-Ching Chen, and Wen-Yin Chen. The design narrative in design learning: Adjusting the inertia of attention and enhancing design integrity. *The Design Journal*, 26(4): 519–535, 2023.

Cas W Coopmans and Neil Cohn. An electrophysiological investigation of co-referential processes in visual narrative comprehension. *Neuropsychologia*, 172:108253, 2022.

Michael F Dahlstrom. Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the national academy of sciences*, 111(supplement_4):13614–13620, 2014.

Michael F Dahlstrom and Shirley S Ho. Ethical considerations of using narrative to communicate science. *Science Communication*, 34(5):592–617, 2012.

Matthew Z Dudley, Gordon K Squires, Tracy M Petroske, Sandra Dawson, and Janesse Brewer. The use of narrative in science and health communication: a scoping review. *Patient Education and Counseling*, page 107752, 2023.

Alison Engel, Kathryn Lucido, and Kyla Cook. Rethinking narrative: Leveraging storytelling for science learning. *Childhood Education*, 94(6):4–12, 2018.

Walter R Fisher. *Human communication as narration: Toward a philosophy of reason, value, and action*. University of South Carolina Press, 2021.

Mary L Gick and Keith J Holyoak. Analogical problem solving. *Cognitive psychology*, 12(3): 306–355, 1980.

Jonathan Gottschall. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt, 2012.

Philip N Hineline. Narrative: Why it's important, and how it works. *Perspectives on Behavior Science*, 41:471–501, 2018.

Keith J Holyoak and Paul Thagard. A computational model of analogical problem solving. *Similarity and analogical reasoning*, 242266, 1989.

David H Jonassen and Julian Hernandez-Serrano. Case-based reasoning and instructional design: Using stories to support problem solving. *Educational technology research and development*, 50 (2):65–77, 2002.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL `https://aclanthology.org/2023.acl-long.291`.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL `https://aclanthology.org/P17-1015`.

Susana Martinez-Conde and Stephen L Macknik. Finding the plot in science storytelling in hopes of enhancing science communication. *Proceedings of the National Academy of Sciences*, 114(31): 8127–8129, 2017.

Areej Mawasi, Peter Nagy, and Ruth Wylie. Systematic literature review on narrative-based learning in educational technology learning environments (2007-2017). 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Aquiles Negrete and Cecilia Lartigue. Learning from education to communicate science as a good story. *Endeavour*, 28(3):120–124, 2004.

Stephen P Norris, Sandra M Guilbert, Martha L Smith, Shahram Hakimelahi, and Linda M Phillips. A theoretical framework for narrative explanation in science. *Science education*, 89(4):535–563, 2005.

Corinna Oschatz and Caroline Marker. Long-term persuasive effects in narrative communication research: A meta-analysis. *Journal of Communication*, 70(4):473–496, 2020.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.294. URL `https://aclanthology.org/2023.acl-long.294`.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Jonathan P Rowe, Lucy R Shores, Bradford W Mott, and James C Lester. Integrating learning and engagement in narrative-centered learning environments. In *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part II 10*, pages 166–177. Springer, 2010.

Dario D Salvucci and John R Anderson. Integrating analogical mapping and general problem solving: the path-mapping theory. *Cognitive Science*, 25(1):67–110, 2001.

Lila San Roque, Lauren Gawne, Darja Hoenigman, Julia Miller, Stef Spronck, Alan Rumsey, Alice Carroll, and Nicholas Evans. Getting the story straight: Language fieldwork using a narrative problem-solving task. 2012.

Anthony J Sanford and Catherine Emmott. *Mind, brain and narrative*. Cambridge University Press, 2012.

Joanna Szurmak and Mindy Thuna. Tell me a story: The use of narrative as tool for instruction. 2013.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022. URL `https://api.semanticscholar.org/CorpusID:247595263`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Daniel T Willingham. Ask the cognitive scientist the privileged status of story. *American Educator*, 28:43–45, 2004.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*, 2024.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*, 2023.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Paul J Zak. Why inspiring stories make us react: The neuroscience of narrative. In *Cerebrum: the Dana forum on brain science*, volume 2015. Dana Foundation, 2015.

A Story of Thought (SoT) vs. Chain of Thought (CoT)

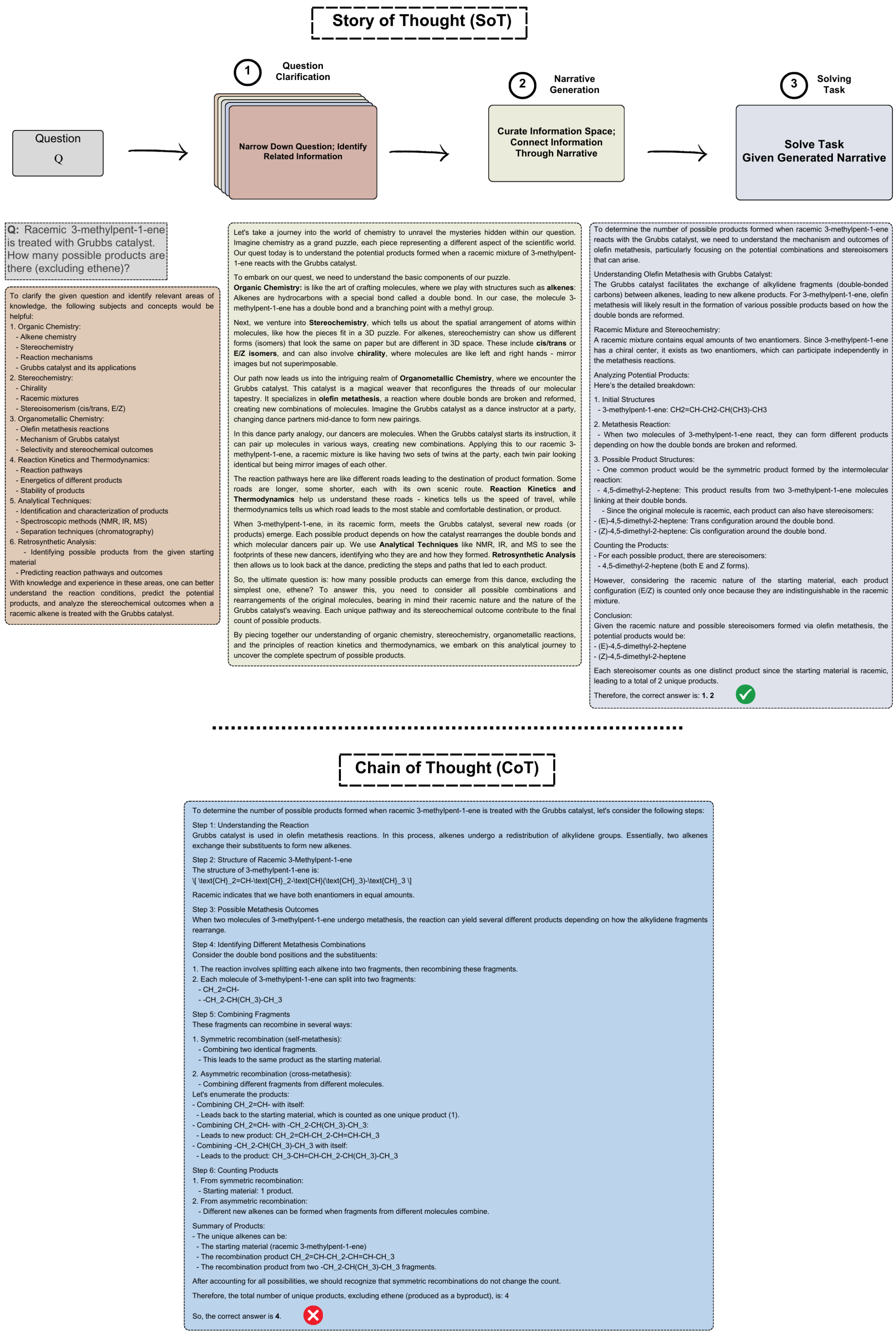

Figure 4: An actual example of SoT.

B Designed Prompts

B.1 Step 1: Question Clarification

```
You are an explorer who wants to identify and collect different related and specialized subject areas to clarify the question. Your goal is to narrow down the question and provide relevant areas of knowledge and experience you have that help clarify the question mentioned below. You should not answer the question.

<question>
```

B.2 Step 2: Narrative Generation

```
You are an expert in narrative-based explanations for science communication. Your goal is to clarify the following question in a narrative way through the interconnected information provided below to enable a non-expert to comprehend the question in a more coherent and contextually rich manner. You should not answer the question.

Make sure to use all of these narrative techniques when clarifying the question through the interconnected information: Progressive Disclosure, Branching, Analogy, Analogical Reasoning, and Metaphor.

<question>

<generated information in the previous step>
```

B.3 Step 3: Solving Task

```
You are an expert in analyzing narrative-based explanations for solving tasks. Please answer the following question based on the following narrative-based clarification:

<question>

Options:
<options>

<generated narrative in the previous step>
```

B.4 Analyzing Generated Narratives

```
You are an expert in analyzing narrative-based explanations for science communication. Your goal is to find out which narrative techniques have been used in the following narrative-based explanation.

Label the narrative-based explanation using the following narrative-based techniques:
1. Progressive Disclosure
2. Branching
3. Analogy
4. Analogical Reasoning
5. Metaphor

<generated narrative>
```