# Can Stories Help LLMs Reason? Curating Information Space Through Narrative

**Vahid Sadiri Javadi** [Ψ] **Johanne R. Trippas** [✠] **Yash Kumar Lal** [ψ] **Lucie Flek** [Ψ]
[Ψ] University of Bonn, [✠] RMIT University, [ψ] Stony Brook University
{vahid.sadirij, lflek}@uni-bonn.de,
j.trippas@rmit.edu.au, ylal@cs.stonybrook.edu

## Abstract

Narratives are widely recognized as a powerful tool for structuring information and facilitating comprehension of complex ideas in various domains such as science communication. This paper explores whether generating narratives can serve "as a specialized mode of thinking" that improves the reasoning abilities of Large Language Models (LLMs). We introduce **Story of Thought (SoT)**, a novel prompt-driven reasoning framework that guides LLMs to construct narratives around the problem statement to solve the task more effectively. SoT enables LLMs to integrate narrative techniques such as *metaphor* and *analogy* into their reasoning process. Our experiments show that SoT significantly improves the LLMs' problem-solving abilities on various tasks including physics, chemistry, and biology in both JEEBench and GPQA (e.g., SoT resulted in 13% improvement compared to CoT when using GPT-4). To validate LLM-based evaluation for generated narratives, we conduct a human annotation of the narrative techniques used by LLMs. Our results show strong inter-annotator agreement between Llama 3 70B and human annotators. This work brings LLM reasoning closer to human cognitive processes by mirroring mechanisms such as analogical problem-solving, which are central to how humans understand and process complex ideas.

## 1 Introduction

Humans employ two fundamental modes of thought: the *logico-scientific mode* which relies on formal logic and a mathematical system of description to derive conclusions, whereas the *narrative mode* organizes information into structured stories, making sense of complex ideas through causality (Bruner, 1991). Literature on human cognition has extensively explored how the human brain processes narratives, highlighting humans' exceptional ability to understand and reason through them (Hineline, 2018; Armstrong, 2020; Sanford and Emmott, 2012). A narrative-driven explanation can enhance the comprehension and retention of complex subjects compared to a simple listing of objective information (Fisher, 2021; Abbott, 2020; Gottschall, 2012). Storytelling effectively structures information in science communication (Dahlstrom, 2014; Norris et al., 2005; Martinez-Conde and Macknik, 2017) and education (Engel et al., 2018; Negrete and Lartigue, 2004), revealing relationships and contextual nuances (Zak, 2015). Figure 1 shows an example of the *narrative approach* that contextualizes facts within a daily life scenario (story) with a planned structure, allowing for the use of techniques such as analogy or progressive disclosure, while the *logico-scientific approach* conveys information in a concise in-domain manner.

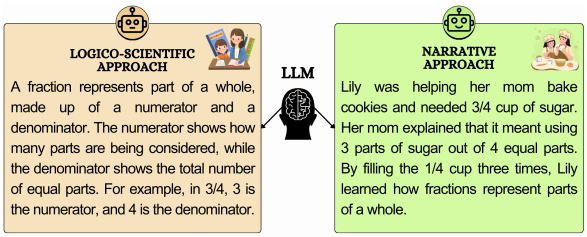


Figure 1: A comparison of narrative and logico-scientific explanations for the concept of fractions. The narrative approach places the concepts into a progressively rolled-out story, while the logico-scientific approach presents the information objectively.

To date, one of the ways the reasoning process in large language models (LLMs) has been enhanced is through prompting techniques that guide them to break tasks into smaller subtasks such as Chain-of-Thought (CoT) (Wei et al., 2022) and its more recent adaptations (Xia et al., 2024). The strategies of constructing natural language rationales (Ling et al., 2017), in the CoT context play a vital role in LLM prompting (Ye and Durrett, 2022; Min et al.,



*Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 92–107

2022; Wang et al., 2022; Li et al., 2023). However, LLMs still 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).

In this work, we show that generating narratives around the problem statement enhances the LLMs' reasoning ability. Our method integrates narrative techniques such as *analogy* into the reasoning process with the aim of combining their effectiveness in explaining abstract concepts with their ability to organize information flow coherently. Therefore, we address two main research questions:

**RQ1:** How to leverage LLMs to generate narratives around problem statements to facilitate comprehension and reasoning?

**RQ2:** Can incorporating narratives into the reasoning process improve model performance on complex problem-solving tasks?

We make the following contributions: *(i)* We introduce a novel method, **Story of Thought (SoT)**, that aids LLMs in identifying and arranging relevant information for solving complex tasks by incorporating narrative structures into the reasoning process, *(ii)* We evaluate the effectiveness of SoT on GPQA and JEEBench datasets of complex problems, showing superior performance to existing prompting techniques with SotA models, and *(iii)* We analyze the impact of narrative techniques to generate narrative-based explanations and investigate why they improve LLMs' reasoning abilities.

## 2 Related Work

Bruner (1991) posit that narratives are a fundamental mode of human thought, allowing individuals to convey complex concepts in a more understandable manner. Presenting information through narratives can enhance learning and memory, promote engagement and motivation (Willingham, 2004; Chen et al., 2023). The development of narrative-based educational strategies (Bower and Clark, 1969; Mawasi et al., 2020; Norris et al., 2005) paved the way for using them as a framework for organizing information for problem solving. The use of narratives can break down complex problems into sub-problems, providing a step-by-step approach to answering a question (Szurmak and Thuna, 2013). Sadiri Javadi et al. (2024) use different narratives techniques to satisfy diverse requirements for conversational information-seeking systems.

There are a plethora of datasets focusing on answering questions about given contexts. Reading comprehension datasets (Khashabi et al., 2018; Welbl et al., 2018; Williams et al., 2018; Mihaylov et al., 2018) explicitly evaluate a system's ability to answer questions that need information from multiple sentences in a passage. NarrativeQA (Kočiský et al., 2018) provides a dataset of 1,567 narratives and associated QA pairs as written by human annotators. ROCStories (Mostafazadeh et al., 2016) is a collection of 5 sentence short stories over which numerous datasets such as TellMeWhy (Lal et al., 2021) have been built to facilitate answering questions about narratives. However, none of these datasets use narratives as a tool of understanding, or relate to problem solving.

Problem solving datasets focus on mathematics, physics or other scientific domains. GSM8K (Cobbe et al., 2021) is a dataset of 8.5K high quality linguistically diverse grade school math word problems created by human problem writers. SciQ (Welbl et al., 2017) is built using a novel method for obtaining high-quality, domain-targeted multiple choice questions from crowd workers, and contains 13.7K multiple choice science exam questions. ScienceQA (Lu et al., 2022) adds multimodal context to collected elementary and high school science questions. While there has been rapid progress on these tasks, prior work has not integrated educational strategies such as narratives to tackle them, a setting which is likely to be used in the real world. MedMCQA (Pal et al., 2022) contains MCQ questions designed to address real-world medical entrance exam questions. Such datasets have been used extensively as yardsticks to measure the progress of NLP techniques.

The strength of modern LLMs, coupled with the paradigm of prompting, has driven up performance on problem solving tasks. In-context learning through few-shot examples has been used to teach LLMs about new tasks using a small number of examples. Chain of thought prompting (Wei et al., 2022) nudges LLMs to generate intermediate steps to mimic an explicit reasoning process before answering a question. Similarly, Tree of Thoughts (ToT) (Yao et al., 2023) and Graph of Thoughts (GoT) (Besta et al., 2024) induce intermediate reasoning structures, trees and graphs respectively, to decide on an answer. However, despite the fact that narratives have been used as a way to simplify problems, they have never been explored to improve the problem solving abilities of LLMs.

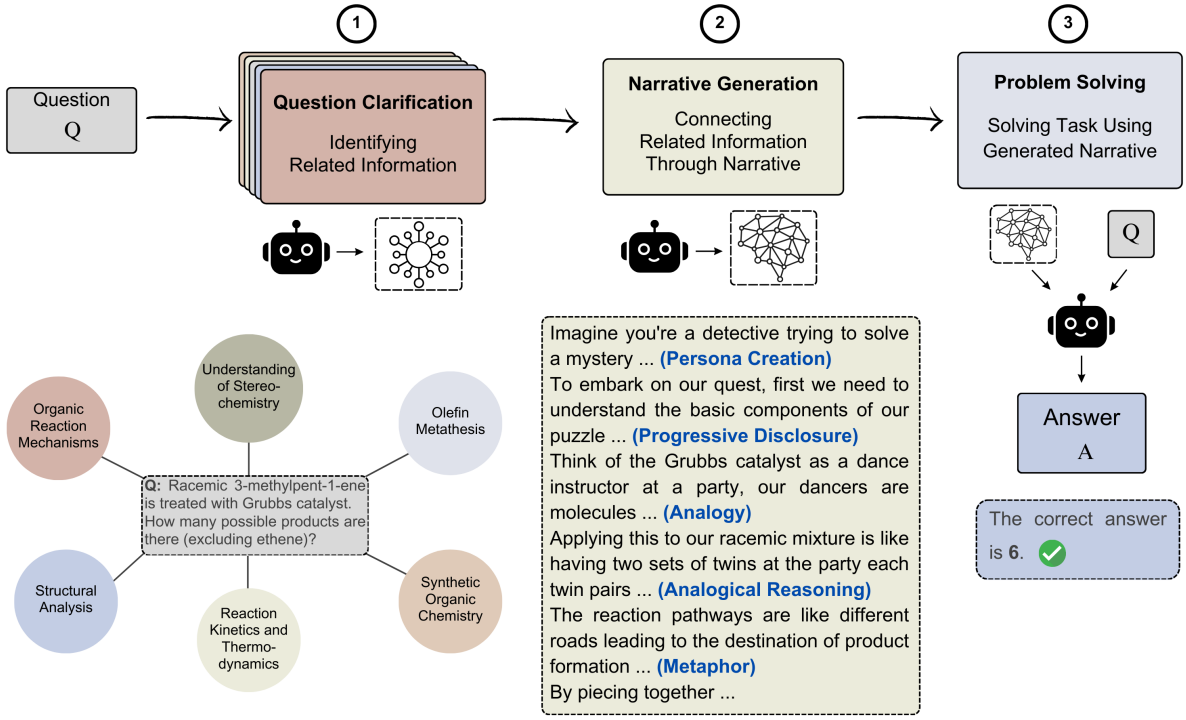


Figure 2: A high-level overview of **Story of Thought** (SoT), consisting of three steps (top): ① Question Clarification (See Section 3.1), ② Narrative Generation (See Section 3.2), ③ Problem Solving (See Section 3.3) and an actual example of LLM output (bottom) in each step for the GPQA task. See Appendix C for prompts for each step. 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 G for an example of SoT.

## 3 Methodology: Story of Thought

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: *(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)* **Problem Solving** (i.e., leveraging generated narratives by LLMs to solve the tasks (Section 3.3)). We describe the exact prompts used in each step in Appendix C.

### 3.1 Step 1: Question Clarification

In the first step, we use the LLM's ability to explore and clarify the question. 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.

### 3.2 Step 2: Narrative Generation

The second step involves generating detailed narratives based on the breakdown and clarification performed in Step 1 (question clarification). These narratives provide a structured context for the questions to enhance the LLM's understanding, reasoning, and problem-solving abilities. Sadiri Javadi et al. (2024) discuss different narrative techniques required in conversational information-seeking systems. We integrate the below subset of these techniques into our prompt and task LLMs to generate a narrative, based on the information from Step 1:

1. **Progressive Disclosure (PD)**: Reveals information gradually, guiding the LLM step-by-step through the problem-solving process.
2. **Branching (BR)**: Explores different paths or approaches to understanding the problem by pro-

viding multiple perspectives.

3. **Analogy (AN)**: Uses comparisons to familiar concepts or situations to make abstract components more understandable.
4. **Analogical Reasoning (AR)**: Facilitates understanding by reasoning through similarities between the problem and known situations.
5. **Metaphor (ME)**: Simplifies complex ideas through metaphorical representation.

The selection of these narrative techniques was grounded in cognitive science and educational psychology principles, which emphasize their effectiveness in enhancing comprehension and reasoning. Analogical reasoning and analogy were chosen for their proven ability to map complex, abstract problems onto familiar concepts, facilitating understanding, problem-solving (Gentner and Smith, 2013), and analogical reasoning (Holyoak and Lu, 2021). Metaphors, similarly, reveal how humans transfer knowledge between domains through structural mapping (Chiu, 2000; Thibodeau and Boroditsky, 2011). Branching aligns with decision-making frameworks that explore alternative paths to solutions, mirroring human problem-solving strategies (Yao et al., 2024). Progressive disclosure — rooted in cognitive load theory (Sweller, 1988) — reflects how humans incrementally process and integrate new information to manage cognitive load and maintain focus (Chandler and Sweller, 1991).

### 3.3 Step 3: Problem Solving

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.

## 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, math, and chemistry problem-solving. In particular, we utilize the **GPQA** (Diamond set) (Rein et al., 2024) and **JEEBench** (Arora et al., 2023). GPQA is a Graduate-level Problem-solving QA dataset that comprises expert-crafted multiple-choice questions. We use the Diamond set of GPQA, which contains 198 questions written by domain experts in biology, physics, and chemistry of high quality and difficulty. JEEBench contains 515 challenging pre-engineering mathematics, physics and chemistry problems from the highly competitive IIT JEE-Advanced exam.

### 4.2 Benchmarking 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 & Llam-3-70B (instruction-tuned versions)
2. **Mistral**: Mistral 7B & Mixtral 8x7B
3. **OpenAI**: GPT-3.5-turbo & GPT-4-turbo
4. **Microsoft**: Phi-3-Medium & Phi-3-Mini

These models were selected to cover a wide spectrum of capabilities, sizes and families, enabling a comprehensive evaluation of their strengths and limitations. More details on the implementation can be found in Appendix B.

### 4.3 Methods Studied

We compared our proposed approach against several prompting techniques:

**Zero-shot Prompting**: LLMs are prompted to solve tasks based solely on their pre-trained knowledge without any labeled examples or explicit guidance.

**Zero-shot CoT** (Wei et al., 2022): We prompt the LLM to explicitly reason through the steps required to arrive at an answer (i.e., "think step by step and answer the question."). This aims to improve the model's ability to solve complex problems by breaking them down into smaller, more manageable steps.

**Tree of Thoughts** (Yao et al., 2023): This method systematically explores multiple reasoning paths instead of a single linear progression. In ToT, a tree-structured solution to a problem is generated by breaking it down into sub-problems. This enables the model to consider a broader set of potential solutions by evaluating each branch for correctness before proceeding further.

**Graph of Thoughts** (Besta et al., 2024): This technique extends the Tree of Thoughts (ToT) approach by allowing for a more flexible and non-hierarchical representation of problem-solving steps. The reasoning steps are treated as nodes, and the connections between them are edges that represent logical relationships or dependencies.

| 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 |
| **Tree of Thoughts** | 34.34 | 43.43 | 29.79 | 32.82 | 24.24 | 42.42 | 18.68 | 31.81 |
| **Graph of Thoughts** | 33.83 | 43.43 | 28.78 | 30.30 | 23.23 | 40.90 | 19.69 | 28.78 |
| **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 |

Table 1: On GPQA (Diamond set), Story of Thought (SoT) consistently outperforms other techniques. We present the performance (QA accuracy) of different methods with various LLMs on GPQA Diamond set.

**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 (i.e., utilizing narrative in reasoning), we prompt LLMs to solve the task based solely on the generated knowledge from Step 1 (described in Section 3.1). This allows us to compare the model's 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

Our proposed SoT approach that incorporates narrative structures improves over almost all previous prompting approaches across two different problem-solving datasets. This highlights the potential of using narratives to improve the ability of LLMs to understand and reason about the given information in various intensive reasoning tasks.

### 5.1 Performance on GPQA

Results on GPQA (Diamond set) are presented in Table 1. For this task, SoT is the best method to use with six of eight models. The open-source Llama 3 70B records the highest accuracy 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. Furthermore, the GPT-4 model shows the most notable improvement in accuracy with SoT, compared to the 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).[1] Interestingly, all reasoning strategies lead to an accuracy drop for the 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 and study it further in §6.1. We note that, on average, models improve the most on biology problems when using SoT. See Appendix D for subject-wise performance evaluation.

### 5.2 Performance on JEEBench

Table 2 presents detailed experimental results on JEEBench. Our proposed method (SoT) consistently improves the performance of seven out of the eight LLMs. Using SoT, Llama 3 70B performance surpasses even the GPT models. It obtains the highest scores in all subjects and question types (Except Single-Correct), with an overall aggregate score of 0.453. This is a significant improvement on the previous SotA, which was a strong GPT-4 model used with both CoT and Self-Consistency. Across models, the results highlight the effectiveness of Story of Thought (SoT) in enhancing model performance on complex, multi-disciplinary benchmarks like JEEBench, setting new SotA results in several categories. The improvements are particularly notable in the subject categories and question types where the other methods struggle.

[1] We also find that Llama 3 70B with SoT outperforms zero-shot o1-preview which uses CoT style reasoning internally.

| | Chemistry | Mathematics | Physics | Integer | Single-Correct | Multi-Correct | Numeric | Total |
|---|---|---|---|---|---|---|---|---|
| GPT-4+CoT+SC@8* | 0.463 | 0.308 | 0.449 | 0.293 | **0.618** | 0.410 | 0.234 | 0.389 |
| Llama 3 8B | 0.143 | 0.082 | 0.089 | 0.061 | 0.127 | 0.148 | 0.044 | 0.102 |
| Llama 3 8B+CoT | 0.127 | 0.101 | 0.116 | **0.11** | 0.145 | 0.149 | 0.036 | 0.112 |
| Ours: Llama 3 8B+SoT | **0.154** | **0.195** | **0.172** | 0.072 | **0.259** | **0.324** | 0.028 | **0.173** |
| Llama 3 70B | 0.324 | 0.189 | 0.274 | 0.171 | 0.345 | 0.316 | 0.131 | 0.25 |
| Llama 3 70B+CoT | 0.264 | 0.228 | 0.268 | 0.159 | 0.291 | 0.317 | 0.175 | 0.249 |
| Ours: Llama 3 70B+SoT | **0.554** | **0.329** | **0.471** | **0.446** | **0.42** | **0.485** | **0.462** | **0.453** |
| Mistral 7B | 0.119 | 0.079 | 0.091 | 0.049 | 0.109 | 0.159 | 0.022 | 0.094 |
| Mistral 7B+CoT | 0.106 | 0.123 | 0.059 | 0.073 | 0.118 | 0.165 | 0.022 | 0.102 |
| Ours: Mistral 7B+SoT | **0.2** | **0.177** | **0.201** | **0.11** | **0.245** | **0.224** | **0.146** | **0.19** |
| Mixtral 8x7B | 0.22 | 0.151 | 0.167 | 0.122 | 0.218 | 0.261 | 0.058 | 0.176 |
| Mixtral 8x7B+CoT | 0.237 | 0.142 | 0.152 | 0.061 | 0.209 | 0.27 | 0.08 | 0.173 |
| Ours: Mixtral 8x7B+SoT | **0.253** | **0.251** | **0.274** | **0.268** | **0.309** | 0.277 | **0.182** | **0.257** |
| ChatGPT 3.5 | **0.228** | **0.146** | 0.173 | 0.073 | **0.318** | **0.249** | 0.029 | **0.177** |
| ChatGPT 3.5+CoT | 0.17 | 0.111 | 0.167 | 0.11 | 0.173 | 0.206 | 0.051 | 0.142 |
| Ours: ChatGPT 3.5+SoT | 0.189 | 0.128 | **0.189** | 0.073 | 0.291 | 0.204 | 0.051 | 0.161 |
| GPT 4 | 0.423 | 0.212 | 0.352 | 0.207 | 0.455 | 0.383 | 0.153 | 0.309 |
| GPT 4+CoT | 0.468 | 0.280 | 0.335 | 0.256 | **0.473** | 0.448 | 0.175 | 0.350 |
| Ours: GPT 4+SoT | **0.535** | **0.294** | **0.413** | **0.378** | 0.4 | **0.453** | **0.321** | **0.395** |
| Phi-3 Mini | **0.256** | 0.12 | **0.199** | 0.146 | 0.255 | 0.224 | 0.08 | 0.18 |
| Phi-3 Mini+CoT | **0.256** | 0.137 | 0.171 | 0.134 | 0.209 | 0.216 | **0.139** | 0.181 |
| Ours: Phi-3 Mini+SoT | 0.224 | **0.209** | 0.181 | **0.183** | **0.282** | **0.234** | 0.124 | **0.207** |
| Phi-3 Medium | 0.298 | 0.193 | 0.165 | 0.146 | 0.255 | 0.286 | 0.139 | 0.218 |
| Phi-3 Medium+CoT | 0.253 | 0.195 | 0.199 | 0.171 | 0.236 | 0.274 | 0.139 | 0.214 |
| Ours: Phi-3 Medium+SoT | **0.279** | **0.203** | **0.224** | **0.232** | **0.273** | **0.263** | **0.153** | **0.231** |

Table 2: On JEEBench, Story of Thought (SoT) outperforms previous SOTA as well as other methods. We present the aggregate score by subject as well as question type and present the overall aggregate score. The best overall scores are highlighted in **blue** while the best score by method for a model is in **bold**. * reported in (Arora et al., 2023).

## 6 Analysis of SoT Aspects

### 6.1 Role of the Narrative Quality/Choice

The choice of *narrator* model (i.e., the model that generates narratives) can impact the problem-solving resuls. 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 3.

| 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↓) |

Table 3: Applying generated narratives by open-source models to Microsoft models to solve the tasks.

We observe that the **narratives** generated by most 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. Larger models generate narratives that break down problems to make them more easily solvable. Unsurprisingly, there is larger room for improving the problem solving abilities of smaller models.

### 6.2 Impact of Narrative Techniques

To measure the impact of each narrative technique, 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 5 indicate that **employing any single narrative elements at a time is notably less effective at boosting QA accuracy than utilizing a combination of these simultaneously**. For both Llama models, the decrease in accuracy is comparably smaller (-3.0% to -5.6%) when using only the analogical com-

| 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 |

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

| 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↓) |
| **None** | 38.38 (-5.05↓) | 45.45 (-5.56↓) |
| **All** | 43.43 | 51.01 |

Table 5: Comparing accuracy when using a single narrative technique or no narrative technique (None). Values in parentheses represent the decrease in accuracy percentage points compared to a combination of multiple narrative techniques simultaneously (shown in Table 1).

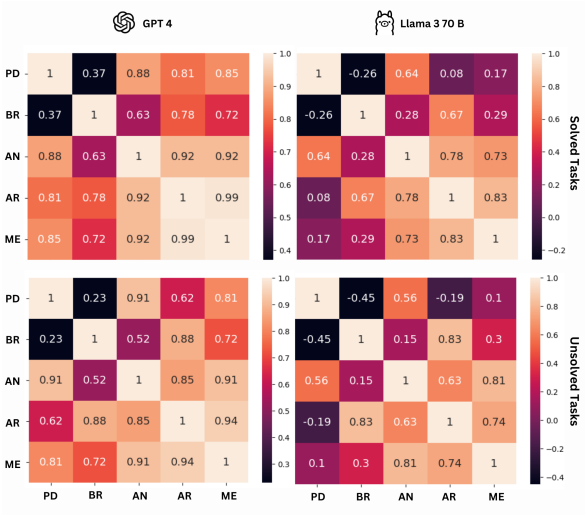


Figure 3: Correlation coefficients among all narrative elements (**PD** = Progressive Disclosure, **BR** = Branching, **AN** = Analogy, **AR** = Analogical Reasoning, **ME** = Metaphor) used in the SoT approach for GPT 4 and Llama 3 70B in solved and unsolved tasks.

ponents 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%). The *None* condition, where no narrative technique is mentioned in the prompt, results in an accuracy drop (-5.0% to -5.6%). This makes it difficult to determine how the narrative techniques relate to each other. We study this going forward.

### 6.3 Analyzing Generated Narratives

To gain deeper insights into the generated narratives, we prompt Llama 3 70B to annotate the number of times each narrative technique appears (i.e., the number of occurrences) in each generated narrative across all models used in our experiments. We can better interpret how the model executed the narrative generation prompt, by asking it to label if and where the mentioned techniques are used in the generated narrative. A proportion of the narrative techniques and their correlation can provide us with a better picture of LLM's interpretation of the instruction as well. The instructions can be found in Appendix C. We aim to uncover patterns and variations in the use of narrative techniques across different LLMs. Table 4 compares 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), LLMs generate narrative using all 5 narrative techniques. However, as Table 4 indicates, not all elements were employed equally. This 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 tech-

| Similarity Metric | BertScore | | ROUGE-L | | BLEU | |
|---|---|---|---|---|---|---|
| | SoT Reasoning | CoT Reasoning | SoT Reasoning | CoT Reasoning | SoT Reasoning | CoT Reasoning |
| **Llama 3 8B** | **0.28** | 0.06 | **0.19** | 0.11 | **6.57** | 0.19 |
| **Llama 3 70B** | **0.3** | 0.04 | **0.2** | 0.1 | **8.18** | 0.06 |
| **Mistral 7B** | 0.27 | **0.33** | 0.18 | **0.2** | **8.12** | 4.65 |
| **Mixtral 8x7B** | 0.3 | **0.34** | 0.19 | **0.21** | **8.92** | 8.14 |
| **ChatGPT 3.5** | **0.3** | 0.24 | **0.19** | 0.16 | 6.1 | **6.07** |
| **GPT 4** | 0.31 | **0.34** | 0.19 | **0.2** | **8.84** | 6.73 |
| **Phi-3 Mini** | 0.27 | **0.31** | 0.17 | **0.19** | **6.54** | 6.36 |
| **Phi-3 Medium** | 0.3 | **0.35** | 0.2 | **0.21** | 7.13 | **8.4** |

Table 6: Comparison of generated Story of Thought (SoT) and Chain of Thought (CoT) reasoning with Human Explanations on the GPQA (Diamond set) using BERTScore, ROUGE-L, and BLEU metrics across various large language models. Bold values indicate the reasoning approach that is more similar to human explanations for each model and metric pair.

niques compared to their smaller counterparts. Furthermore, 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.

### 6.4 Human Evaluation

To assess the reliability of the LLM-based annotation method (described in Section 6.3), we conduct a human evaluation of narrative techniques used by LLMs in generated narratives. We provide 3 annotators with 15 narratives generated by 8 different models, resulting in a total of 120 narratives. Annotators were instructed to identify and count how many times each narrative technique appeared in each narrative. The aggregated annotations were then analyzed using the *Krippendorff Alpha Coefficient* to assess inter-annotator agreement.

The average agreement score across all techniques was 0.72, indicating **strong inter-annotator agreement, with Llama 3 70B aligning closely with human annotators, validating the use of LLM-based evaluation for assessing narrative techniques**. While annotators show the highest agreement in *Branching* and *Analogy*, with average scores of 0.75 and 0.79, respectively, they have in *Metaphor* and *Analogical Reasoning* lower agreement scores (0.69 and 0.68). The extended results can be found in Appendix F.

### 6.5 Analyzing SoT Reasoning

Table 6 compares the similarity of SoT and CoT reasoning outputs to human explanations for different language models on the GPQA (Diamond set) dataset, using BertScore, ROUGE-L, and BLEU.

The differences between ROUGE-L values are insignificant and do not display any clear trends. However, according to BLEU scores, using SoT results in explanations closer to humans and the differences are more pronounced. As per BertScore Llama 3 models' explanations are more similar to human ones when using SoT reasoning across all three metrics. However, Mistral models, GPT-4, and Phi-3 Mini generate explanations more similar to human explanations when using CoT reasoning across all metrics. The semantic similarity of narratives generated by Llama 3 70B to human explanations combined with their effect of improving smaller models indicates that these narratives present information about the problems in a simplified manner.

## 7 Conclusion

Inspired by findings from human cognitive processes explored in didactics research, in this work, we propose to use narratives in LLMs prompting. We present strong evidence on public benchmark datasets that narratives 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.

## Limitations

**Dataset limitations.** So far, we used only GPQA and JEEBench problems 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.** 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 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.

## References


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

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

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

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

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

Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332.

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

Ming Ming Chiu. 2000. Metaphorical reasoning: Origins, uses, development, and interactions in mathematics. *EDUCATION JOURNAL-HONG KONG-CHINESE UNIVERSITY OF HONG KONG-*, 28(1):13–46.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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

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

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

Dedre Gentner and Linsey A Smith. 2013. Analogical learning and reasoning.

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

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

Keith J Holyoak and Hongjing Lu. 2021. Emergence of relational reasoning. *Current Opinion in Behavioral Sciences*, 37:118–124.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

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

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

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

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

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

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Vahid Sadiri Javadi, Johanne R Trippas, and Lucie Flek. 2024. Unveiling information through narrative in conversational information seeking. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.

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

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

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

Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one*, 6(2):e16782.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

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

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

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

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

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

## A Robustness of LLM Predictions

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 7 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.

## B Model Implementation Details

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 (*gpt-3.5-turbo-0125*) and GPT-4-turbo (*gpt-4-turbo-2024-04-09*) models, we use the OpenAI API to generate outputs. Across all models, the results are averages over 5 runs with a temperature of 1.0 and a maximum number of tokens of 8,000. The other parameters are set to their default values. To ensure consistency in the model outputs, we utilized the *Jsonformer* Python package[3], resulting in structured JSON outputs. A t-test was performed, yielding a p-value of 0.032, indicating statistical significance at the conventional 0.05 level.

## C Prompts Used in Story of Thought

We describe the prompts used for each stage in the SoT framework.

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

[3] https://github.com/1rgs/jsonformer

### C.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>
```

### C.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>
```

### C.3 Problem Solving

```
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>
```

### C.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>
```

| 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↓) |

Table 7: 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.

## D Subject-wise Performance Evaluation

Figure 4 presents the subject-wise performance of different models on both GPQA and JEEBench when using SoT across the different problem domains. We observe that, on average, models improve the most on biology problems when using SoT in GPQA. We hypothesize that this is because it is easier to simplify information for graduate-level biology problems that can be used by models to come up with a solution.

In JEEBench, on average, model performance is highest on Chemistry problems when using SoT. This is in contrast to findings on GPQA and could occur due to the difference in the degree of difficulty of problems in the two datasets (graduate level vs high school level). Regardless, improvements on Biology problems are not far behind those for Chemistry.

## E Performance on JEEBench

## F Huamn Evaluation

Table 9 presents the Krippendorff Alpha coefficient measuring inter-annotator agreement between three human annotators and Llama 3 70B across five narrative techniques: Progressive Disclosure (PD), Branching (BR), Analogy (AN), Analogical Reasoning (AR), and Metaphor (ME). Higher values indicate stronger agreement. The overall average agreement of 0.72 shows a strong correlation between LLM-based and human annotations, supporting the validity of the LLM-based evaluation method.

**Annotator Recruitment and Demographics:** We recruited three master's students in computer science, aged between 24 and 27 (one female, two males). The annotators were compensated fairly for their time, ensuring alignment with appropriate compensation standards. Each annotator was provided with 120 narratives and given detailed instructions, including the definitions of each narrative technique as described in Section 3.2, to ensure a consistent evaluation process. Each annotation was performed independently to minimize bias and ensure reliability.

## G Story of Thought (SoT) vs. Chain of Thought (CoT)

| | Chemistry | Mathematics | Physics | Integer | Single-Correct | Multi-Correct | Numeric | Total |
|---|---|---|---|---|---|---|---|---|
| GPT-4+CoT+SC@8* | 0.463 | 0.308 | 0.449 | 0.293 | **0.618** | 0.410 | 0.234 | 0.389 |
| Llama 3 8B | 0.143 | 0.082 | 0.089 | 0.061 | 0.127 | 0.148 | 0.044 | 0.102 |
| Llama 3 8B+CoT | 0.127 | 0.101 | 0.116 | **0.11** | 0.145 | 0.149 | 0.036 | 0.112 |
| Llama 3 8B+Analogical Reasoning (3-shot) | 0.139 | 0.111 | 0.128 | **0.11** | 0.145 | 0.165 | **0.058** | 0.124 |
| Ours: Llama 3 8B+Knowledge Identification | 0.199 | 0.099 | 0.134 | 0.073 | 0.227 | 0.171 | 0.058 | 0.137 |
| Ours: Llama 3 8B+SoT | **0.154** | **0.195** | **0.172** | 0.072 | **0.259** | **0.324** | 0.028 | **0.173** |
| Llama 3 70B | 0.324 | 0.189 | 0.274 | 0.171 | 0.345 | 0.316 | 0.131 | 0.25 |
| Llama 3 70B+CoT | 0.264 | 0.228 | 0.268 | 0.159 | 0.291 | 0.317 | 0.175 | 0.249 |
| Llama 3 70B+Analogical Reasoning (3-shot) | 0.314 | 0.24 | 0.295 | 0.195 | 0.318 | 0.349 | 0.19 | 0.276 |
| Ours: Llama 3 70B+Knowledge Identification | 0.317 | 0.226 | 0.254 | 0.195 | 0.345 | 0.323 | 0.146 | 0.26 |
| Ours: Llama 3 70B+SoT | **0.554** | **0.329** | **0.471** | **0.446** | **0.42** | **0.485** | **0.462** | **0.453** |
| Mistral 7B | 0.119 | 0.079 | 0.091 | 0.049 | 0.109 | 0.159 | 0.022 | 0.094 |
| Mistral 7B+CoT | 0.106 | 0.123 | 0.059 | 0.073 | 0.118 | 0.165 | 0.022 | 0.102 |
| Mistral 7B+Analogical Reasoning (3-shot) | 0.157 | 0.084 | 0.116 | 0.073 | 0.155 | 0.169 | 0.029 | 0.114 |
| Ours: Mistral 7B+Knowledge Identification | 0.109 | 0.055 | 0.063 | 0.037 | 0.091 | 0.117 | 0.022 | 0.073 |
| Ours: Mistral 7B+SoT | **0.2** | **0.177** | **0.201** | **0.11** | **0.245** | **0.224** | **0.146** | **0.19** |
| Mixtral 8x7B | 0.22 | 0.151 | 0.167 | 0.122 | 0.218 | 0.261 | 0.058 | 0.176 |
| Mixtral 8x7B+CoT | 0.237 | 0.142 | 0.152 | 0.061 | 0.209 | 0.27 | 0.08 | 0.173 |
| Mixtral 8x7B+Analogical Reasoning (3-shot) | 0.202 | 0.155 | 0.197 | 0.122 | 0.191 | **0.281** | 0.066 | 0.179 |
| Ours: Mixtral 8x7B+Knowledge Identification | 0.184 | 0.129 | 0.144 | 0.122 | 0.155 | 0.237 | 0.044 | 0.15 |
| Ours: Mixtral 8x7B+SoT | **0.253** | **0.251** | **0.274** | **0.268** | **0.309** | 0.277 | **0.182** | **0.257** |
| ChatGPT 3.5 | **0.228** | **0.146** | 0.173 | 0.073 | **0.318** | **0.249** | 0.029 | **0.177** |
| ChatGPT 3.5+CoT | 0.17 | 0.111 | 0.167 | 0.11 | 0.173 | 0.206 | 0.051 | 0.142 |
| ChatGPT 3.5+Analogical Reasoning (3-shot) | 0.208 | 0.125 | 0.148 | 0.098 | 0.2 | 0.216 | **0.073** | 0.156 |
| Ours: ChatGPT 3.5+Knowledge Identification | 0.155 | 0.141 | 0.167 | **0.122** | 0.209 | 0.188 | **0.073** | 0.151 |
| Ours: ChatGPT 3.5+SoT | 0.189 | 0.128 | **0.189** | 0.073 | 0.291 | 0.204 | 0.051 | 0.161 |
| GPT 4 | 0.423 | 0.212 | 0.352 | 0.207 | 0.455 | 0.383 | 0.153 | 0.309 |
| GPT 4+CoT | 0.468 | 0.280 | 0.335 | 0.256 | **0.473** | 0.448 | 0.175 | 0.350 |
| GPT 4+Analogical Reasoning (3-shot) | 0.479 | 0.286 | 0.396 | 0.305 | 0.4 | 0.43 | 0.307 | 0.371 |
| Ours: GPT 4+Knowledge Identification | 0.481 | 0.287 | 0.386 | 0.293 | 0.373 | 0.452 | 0.314 | 0.373 |
| Ours: GPT 4+SoT | **0.535** | **0.294** | **0.413** | **0.378** | 0.4 | **0.453** | **0.321** | **0.395** |
| Phi-3 Mini | **0.256** | 0.12 | **0.199** | 0.146 | 0.255 | 0.224 | 0.08 | 0.18 |
| Phi-3 Mini+CoT | **0.256** | 0.137 | 0.171 | 0.134 | 0.209 | 0.216 | **0.139** | 0.181 |
| Phi-3 Mini+Analogical Reasoning (3-shot) | 0.205 | 0.159 | 0.195 | 0.146 | 0.264 | 0.218 | 0.088 | 0.182 |
| Ours: Phi-3 Mini+Knowledge Identification | 0.168 | 0.091 | 0.106 | 0.073 | 0.136 | 0.181 | 0.044 | 0.118 |
| Ours: Phi-3 Mini+SoT | 0.224 | **0.209** | 0.181 | **0.183** | **0.282** | **0.234** | 0.124 | **0.207** |
| Phi-3 Medium | 0.298 | 0.193 | 0.165 | 0.146 | 0.255 | 0.286 | 0.139 | 0.218 |
| Phi-3 Medium+CoT | 0.253 | 0.195 | 0.199 | 0.171 | 0.236 | 0.274 | 0.139 | 0.214 |
| Phi-3 Medium+Analogical Reasoning (3-shot) | 0.258 | 0.181 | 0.173 | 0.159 | 0.218 | 0.276 | 0.117 | 0.202 |
| Ours: Phi-3 Medium+Knowledge Identification | 0.288 | 0.163 | 0.205 | 0.207 | 0.236 | 0.235 | 0.161 | 0.211 |
| Ours: Phi-3 Medium+SoT | **0.279** | **0.203** | **0.224** | **0.232** | **0.273** | **0.263** | **0.153** | **0.231** |

Table 8: On JEEBench, Story of Thought (SoT) outperforms previous SOTA as well as other methods. We present the aggregate score by subject as well as question type and present the overall aggregate score. * denotes SOTA results taken from the original paper (Arora et al., 2023).

| Model Name | Narrative Technique | | | | |
|---|---|---|---|---|---|
| | Progressive Disclosure (PD) | Branching (BR) | Analogy (AN) | Analogical Reasoning (AR) | Metaphor (ME) |
| Llama 3 8B | 0.69 | 0.84 | 0.68 | 0.68 | 0.93 |
| Llama 3 70B | 0.77 | 0.68 | 0.82 | 0.61 | 0.73 |
| Mistral 7B | 0.65 | 0.76 | 0.97 | 0.65 | 0.62 |
| Mixtral 8x7B | 0.67 | 0.78 | 0.74 | 0.82 | 0.76 |
| ChatGPT 3.5 | 0.6 | 0.69 | 0.67 | 0.68 | 0.69 |
| GPT 4 | 0.64 | 0.81 | 0.82 | 0.8 | 0.6 |
| Phi-3 Mini | 0.69 | 0.73 | 0.79 | 0.62 | 0.61 |
| Phi-3 Medium | 0.66 | 0.69 | 0.79 | 0.61 | 0.61 |
| Average | 0.67 | 0.75 | 0.79 | 0.68 | 0.69 |

Table 9: Krippendorff Alpha Coefficient for Human and LLM Annotations.

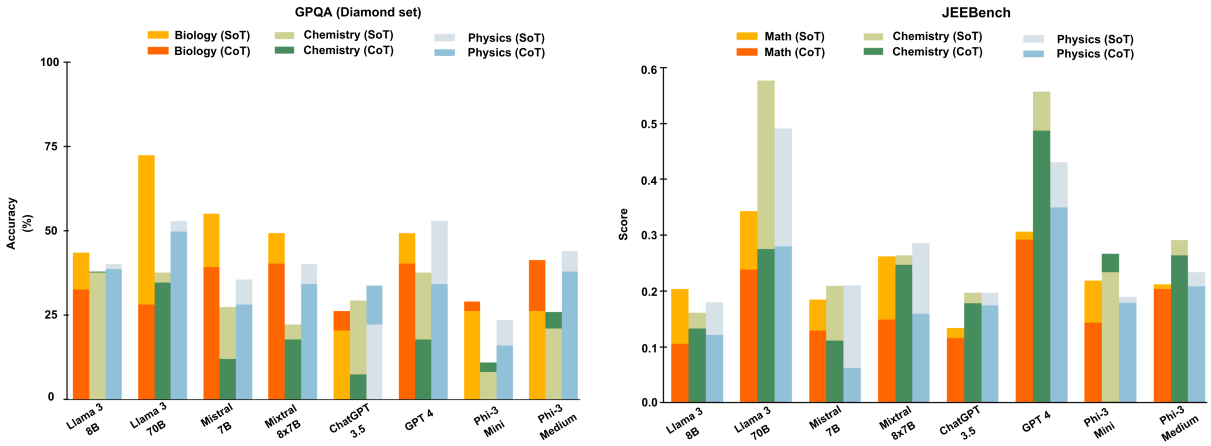


Figure 4: Performance of **Story of Thought** (SoT) on GPQA and JEEBench across various LLMs and domains.

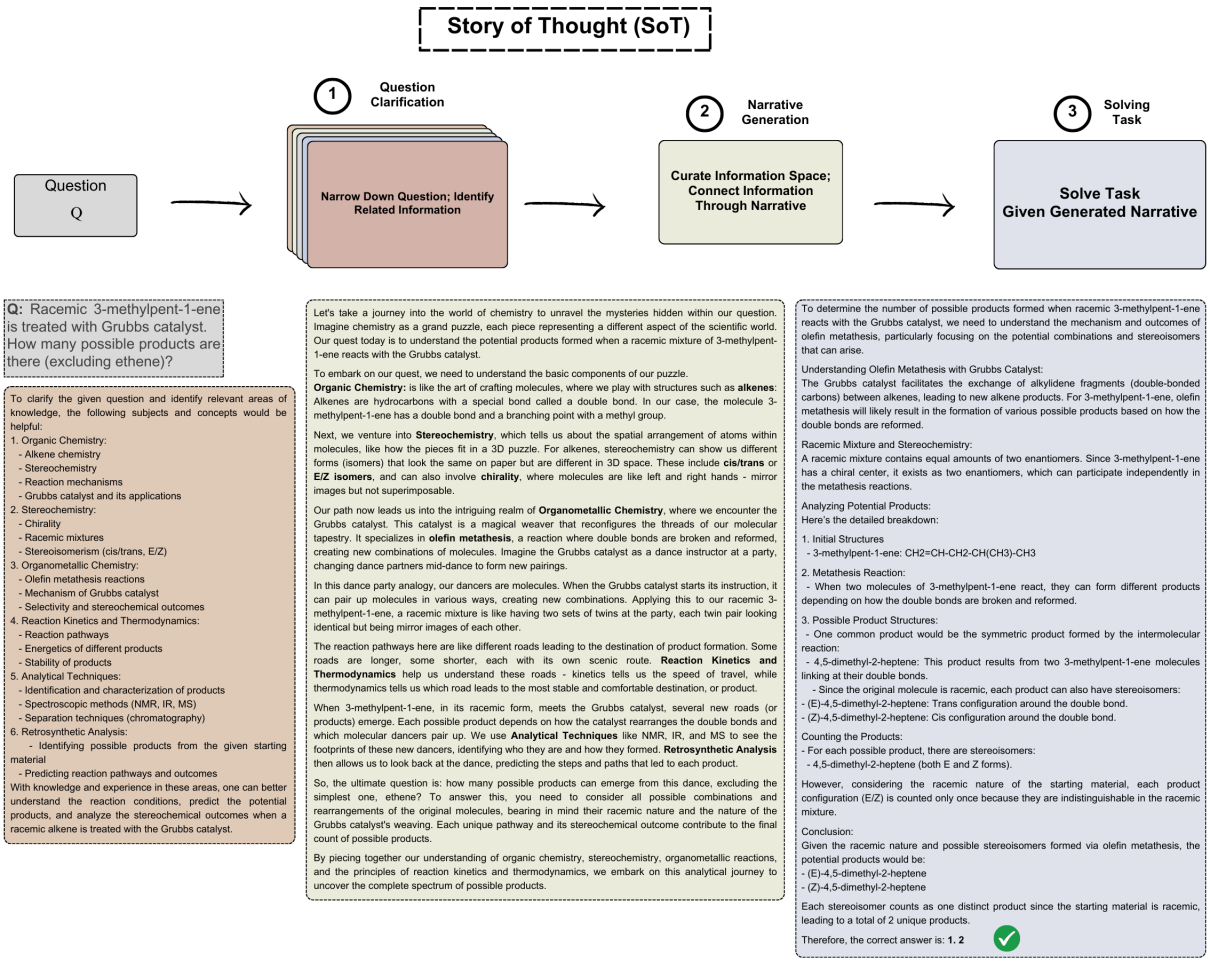


Figure 5: An actual example of SoT.