Can Language Models Serve as Analogy Annotators?

Anonymous ACL submission

Abstract

Conceptual abstraction and analogy-making are crucial for human learning, reasoning, and adapting to unfamiliar domains. Recently, large language models (LLMs) have made the synthesis of analogical data possible, which, however, still heavily relies on extensive human efforts to be annotated. This paper empirically examines the LLMs' capability to annotate story-level analogical data. Specifically, we propose a novel multi-stage progressive reasoning prompt framework A3E (Automated Analogy Annotation Expert), which is based on the structure mapping theory from cognitive psychology and efficiently annotates candidate story pairs across six fine-grained categories. We use A3E to evaluate how well the state-of-the-art LLMs can serve as analogy annotators. Experimental results demonstrate that our proposed A3E achieves an average performance gain of + 73% across a range of prompting baselines and base LLMs. The code and data will be available at https://anonymous.4open.science/r/A3E.

1 Introduction

The ability to abstract concepts and form analogies is fundamental to human learning, reasoning, and the flexible application of knowledge to unfamiliar domains. Analogies are essential for human cognition, and these abilities are equally critical to the development of artificial general intelligence (AGI) that can adapt flexibly and effectively to various domains (Mitchell, 2021).

Recent studies have demonstrated that large language models (LLMs) possess the emergent capability to function as analogical reasoners in a wide spectrum of analogy problems (Webb et al., 2023; Yasunaga et al., 2023). In addition, some works utilized LLMs to generate natural language analogous corpus, thereby leading to a spate of novel research (Jiayang et al., 2023; Sultan et al., 2024).

Although analogy generation holds great promise, a significant challenge lies in the rigorous evaluation and precise annotation of candidate analogies. For example, STORYANALOGY (Jiayang et al., 2023) conducted crowd annotations on Amazon Mechanical Turk (AMT) [1] to evaluate each candidate story pair. To make a proof-of-concept for the generated analogy candidates, ParallelPARC (Sultan et al., 2024) labeled 828 candidate instances on AMT, obtaining 310 gold-set analogy pair paragraphs, for a total cost of $1,804. (Sourati et al., 2023) manually verified all the generated narratives to ensure their quality.

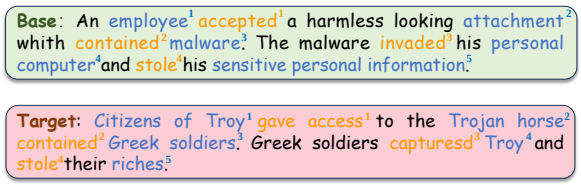

Figure 1: An example of analogical mapping from entities in the base domain to entities in the target domain. The corresponding blue object attributes with the same superscripts can be aligned based on a common relational structure – that is, orange relational verbs with the same superscripts. (Best viewed in color)

These studies underscore the essential role of human annotation in analogy research, but they also reveal its inherent limitation: the process is labor-intensive and costly, imposing a bottleneck for analogy research. As LLMs can generate analogous corpora, can we use them as analogy annotators to effectively inspect the quality of the synthesized analogical data?

We delve into this question at the level of story analogies, which compare entire narratives or coherent sequences of events that involve different entities but similar relations. These analogies enable intelligent agents to understand complex real-world phenomena (Webb et al., 2023) and gain cognitive insights (Bhavya et al., 2023; Ding et al., 2023).

In particular, we evaluated the state-of-the-art LLMs, GPT-4, and LLLAMA-3 on an extended

[1] https://www.mturk.com

set of analogies from (Gentner et al., 1993). We experimentally demonstrate that the most sophisticated LLMs may not possess an emergent capability in analogy annotation tasks by common reasoning prompt methods, such as zero-shot, few-shot CoT (Wei et al., 2022), and instructions like "think step by step" (Kojima et al., 2022). The complexity of the story's analogical mapping process is manifested in the implicit causal relations between events - relations between relations. The higher-order relational analogical mapping plays a crucial role in the construction of the human mind but is lacking in LLMs. To trigger an outbreak of new research with high-quality datasets (e.g., ImageNet (Deng et al., 2009)), we decide to drive the progress of LLMs in analogy annotation.

In this paper, we propose A3E (Automated Analogy Annotation Expert), a multi-stage progressive reasoning prompt framework inspired by Structure Mapping Theory (Gentner, 1983), which pioneers the automated annotation of story-level analogies, leveraging the advanced natural language understanding capabilities of LLMs. Instead of focusing solely on aligning entities and verbs across story pairs, we explicitly align sentences to capture their underlying causal structures. Finally, we categorize the types of analogies into six specific labels.

In summary, our contributions are:

- In a broad survey and empirical analysis, we find that the annotation of story-level analogy corpora is challenging, and even the state-of-the-art LLMs are not up to the task under the guidance of chained reasoning prompts.
- We introduce A3E, a multi-stage progressive reasoning prompt framework for story-level analogy annotation tasks. Experimental results show that our framework achieves an average performance increase of +73% across various prompting baselines and base LLMs.

2 Preliminaries

Gentner's Structure Mapping Theory (SMT) (Gentner, 1983) has become an important theoretical foundation for the study of analogical reasoning in cognitive science (Bach, 2011; Gentner and Maravilla, 2017). According to SMT, the core process in analogy is the analogical *mapping* from entities in base domain $\mathcal{B}$ to entities in target domain $\mathcal{T}$. The mapping process relies on a common relational structure between $\mathcal{B}$ and $\mathcal{T}$, rather than their object attributes. As shown in Figure 1, a personal computer being invaded by malware is analogous to the Trojan Horse incident. In this analogy, there are the following object attributes mapping: *employee* $\rightarrow$ *citizens of Troy*, *attachment* $\rightarrow$ *trojan horse*, *malware* $\rightarrow$ *greek soldiers*, *personal computer* $\rightarrow$ *troy*, and *sensitive personal information* $\rightarrow$ *riches of troy*. It can be seen that the object attributes in the two situations are quite different. However, upon examining the structural relationship mapping: *accept* $\rightarrow$ *give access*, *contains* $\rightarrow$ *contains*, *invade* $\rightarrow$ *capture*, and *steal* $\rightarrow$ *steal*, we can realize that the causal structure of the two domains is similar and reflects the same fact: vulnerabilities will cause serious damage.

We focus on story-level analogies, aiming to classify each analogy pair $(\mathcal{B}, \mathcal{T})$ into one of six specific categories as identified by the LLM annotators. Inspired by SMT and the verified story analogy sources (Gentner et al., 1993), we use the following six labels:

- **Literal Similarity**: Similar to Base in entities (objects and characters), first-order relations (mainly spatial, temporal, and interactional relations), and higher-order relations (chiefly causal relations).
- **True Analogy**: Similar to Base in higher-order relations and many (though not all) first-order relations, dissimilar in entities.
- **False Analogy**: First-order relational match. Similar to Base in first-order relations; dissimilar in entities and higher-order relations.
- **Surface Similar**: Similar to Base in entities and first-order relations but not in higher-order relations.
- **Mere Appearance**: Entities-only match; dissimilar in first-order and higher-order relations.
- **Anomaly**: The entities and relations do not match.

Figure 2 shows a set of analogy samples with one Base and six different Targets to aid in further understanding the concepts of the labels mentioned above.

Base: An employee accepted a harmless looking attachment with contained malware. The malware invaded his personal computer and stole his sensitive personal information.					
Literally Similar	**True Analogy**	**False Analogy**	**Surface Similar**	**Mere Appearance**	**Anomaly**
A worker received a seemingly innocuous attachment that harbored malicious software. The software infiltrated his personal device and pilfered his confidential personal data.	Citizens of Troy gave access to the Trojan horse contained Greek soldiers. Greek soldiers capturesd Troy and stole their riches.	A homeowner went on vacation, leaving their front door unlocked. Burglars took advantage of this opportunity and broke in, stealing their valuable possessions.	An employee opened a harmless-looking large file from a colleague. The file crashed his personal computer, causing loss of his unsaved personal documents.	A worker received a document, containing a hidden puzzle. Intrigued, he spent hours solving the challenging puzzle, enhancing his problem-solving abilities.	Sarah had always been an early riser. She would wake up at 5 a.m. every day without fail, enjoying the tranquility of the early morning hours before starting her day.

Figure 2: An example of analogy samples.

3 Methodology

We develop a method that investigates and validates the abilities of modern LLMs to serve as analogy annotators concerning story analogies. Annotating a pair of story narratives is a reasoning problem. To address this reasoning problem using LLMs, we have formulated it as a multi-step and cascading reasoning generation task, thereby simplifying the annotation task into more manageable sub-tasks. In contrast to traditional CoT-based reasoning methods, each step in our method is more akin to an independent agent, responsible for addressing its specific segment of the reasoning process. This approach brings an additional advantage by reducing the complexity of analogical reasoning to a level accessible to small language models, as it decouples the long dependencies of complex reasoning. In the following sections, we introduce four steps to solve this annotation task: entity analysis (3.1), sentence mapping (3.2), relational alignment (3.3), and analogical conclusion (3.4).

3.1 Entity Analysis

SMT posits that analogical mapping encompasses the similarity of entities, first-order relations, and higher-order relations. Therefore, we commence by analyzing the entity-related facets within the narrative structures of both the Base and Target stories. In our proposed method, we not only employ the LLM to extract similar vocabulary from two narrative stories but also to assess the similarity in terms of background setting, character roles, and responsibilities, as well as plot progression and dynamics.

3.2 Sentence Mapping

A narrative story is generally composed of a series of short sentences. They explicitly present spatial, temporal, and interactive relationships (i.e. first-order relations) and implicitly express causal structural relations (i.e., higher-order relations). Due to temporal and causal dependencies between events, if both Base and Target are considered **True Analogy**, the narrative pair should exhibit a one-to-one sentence mapping in terms of sequence order. Based on this finding, in this step, we instruct the LLM to map sentences with the same structural relationships between the given Base and Target narratives, following the sequence of story presentation. It is worth noting that not all short sentences in the Base or Target can be matched with corresponding sentences. Therefore, for these unmatched sentences, it is sufficient to let the LLM proceed without corresponding matches.

3.3 Relational Alignment

Here, we align the causal structural relations between all short sentence pairs obtained from the previous step. The types of analogy alignment are classified into three groups: *similar*, *dissimilar*, and *irrelevant*. Similar alignment refers to the sentence from Base and its corresponding sentence in Target being able to abstract a common causal relational pattern. Otherwise, it is put into a dissimilar group. For example, "An employee accepted a harmless looking attachment with contained malware" (from Base in Figure 2) and "Citizens of Troy gave access to the Trojan horse contained Greek soldiers" (from True Analogy in Figure 2) are similar, which both involve a harmful object being accepted due to oversight. On the other hand, "The malware invaded his personal computer and stole his sensitive personal information" (from Base in Figure 2) and "Burglars took advantage of this opportunity and broke in, stealing their valuable possessions" (from False Analogy in Figure 2) are considered dissimilar. Although both involve harm to the subjects, the cause in Base is due to active oversight, while

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.27	0.19	0.13	0.22	0.16	0.19	0.10	0.23	0.49	0.34	0.30	0.41
0-shot CoT	0.17	0.23	0.13	0.23	0.19	0.23	0.13	0.24	0.49	0.32	0.29	0.39
1-shot CoT	0.16	0.26	0.19	0.26	0.43	0.31	0.28	0.38	0.46	0.37	0.30	0.44
3-shot CoT	0.19	0.28	0.20	0.28	0.49	0.36	0.33	0.43	0.43	0.36	0.33	0.43
Ours: A3E	**0.34**	**0.37**	**0.34**	**0.44**	**0.63**	**0.59**	**0.61**	**0.71**	**0.57**	**0.54**	**0.54**	**0.64**

Table 1: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, and **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset. Humans performance: Prec: 0.66, Rec: 0.64, F1: 0.64, Acc: 0.63.

in False Analogy, it is due to accidental negligence. In addition, isolated sentences without a match, are grouped as irrelevant.

3.4 Analogical Conclusion

Based on the labels of entities and structural relations obtained from Section 3.1 and Section 3.3, we make the final annotations for each pair of narrative story pairs under the six types according to the following rules (Algorithm 1).

Algorithm 1: Analogy Annotation

Input: background (bool) $\boldsymbol{B}$, role (bool) $\boldsymbol{R}$, plot (bool) $\boldsymbol{P}$, similar group (list) $\boldsymbol{SG}$, dissimilar group (list) $\boldsymbol{DG}$, irrelevant group (list) $\boldsymbol{IG}$.
Output: Label of the given narrative pair $\mathbf{P}(\mathcal{B}, \mathcal{T})$.
if $\boldsymbol{B}$ *is False and* $\boldsymbol{R}$ *is False* **then**
  Entities = *dissimilar*
else
  Entities = *similar*
end
if *len(*$\boldsymbol{SG}$*) > (len(*$\boldsymbol{DG}$*)+len(*$\boldsymbol{IG}$*)/2) or (*$\boldsymbol{B}$ *is True and* $\boldsymbol{R}$ *is True and* $\boldsymbol{P}$ *is True)* **then**
  First-order Relations = *similar*
else
  First-order Relations = *dissimilar*
end
if *len(*$\boldsymbol{DG}$*) > 0 or len(*$\boldsymbol{SG}$*) < len(*$\boldsymbol{IG}$*)/2* **then**
  Higher-order Relations = *dissimilar*
else
  Higher-order Relations = *similar*
end
Annotate analogy for $\mathbf{P}(\mathcal{B}, \mathcal{T})$ based on the definition of analog labels in Section 2.
Return the predicted label of $\mathbf{P}(\mathcal{B}, \mathcal{T})$.

4 Experiments

We experiment with several popular base LLMs: LLAMA3.1(-8B, -70B) (Dubey et al., 2024), and GPT-4o (a flagship version of GPT-4 (OpenAI et al., 2023)) to evaluate the proposed A3E method in story analogy dataset (Gentner et al., 1993). We employ 0-shot, 0-shot Chain-of-Thought (CoT) (Kojima et al., 2022), and few-shot CoT (Wei et al., 2022) (with exemplars K selected from the set 1, 3) as benchmark comparison methods. The performance metrics, including Precision, Recall, F1, and Accuracy, are summarized for each method across different base LLMs in Table 1.

We have the following observations: (1) In terms of story analogy reasoning annotation task, even the state-of-the-art LLMs, have not acquired sufficient emergent ability to find zero-shot solutions. This finding is distinct from the fact observed in Webb et al.'s study (Webb et al., 2023), which simply asked, "Which of Story A and Story B is a better analogy to Story 1?". This might be because making comparisons is relatively easier than providing annotations. (2) The CoT prompting does help with analogical reasoning, and an increase in the number of exemplars could also further improve performance, but the advantage is limited. (3) Our prompting method outperforms all baselines in both LLAMA3.1 and GPT-4o. Although it performs poorly on smaller LLM (8B), it is encouraging that a 70B open-source model will be sufficient to drive the research in analogical reasoning. In addition, it is surprising to note that the A3E combined with LLAMA3.1-70B has reached a level comparable to humans performance.

5 Conclusion

In this paper, we identify the persistent reliance on labor-intensive manual annotation in current research for story-level analogy generation. To address this, we experimentally evaluate the performance of current state-of-the-art LLMs (e.g., GPT4 and LLAMA3) on the analogy annotation task, and propose a novel analogy annotation prompting framework for this task. Experimental results show that our proposed framework significantly outperforms 0-shot CoT and few-shot CoT baselines across different base LLMs. We hope this work will stimulate and contribute to the research on the general reasoning capabilities of LLMs.

6 Limitations

While the proposed A3E demonstrates considerable promise in improving the performance of language models on story-level analogy annotation tasks, it is critical to acknowledge certain limitations, which are important for guiding future extensive research. Since high-quality, professional, and authoritative story-level datasets remain scarce to this day, we were compelled to conduct validation experiments solely on the small and sparse dataset created by cognitive psychologist Gentner (Gentner et al., 1993). This single validation may inherently bias the results in broader scenarios, thereby affecting the wider applicability of the research findings.

On the other hand, although the prompting method we designed is effective compared to other baseline methods on small-scale language models, its accuracy remains below the standard required for practical production applications, Consequently, it also imposes relatively high demands on computational resources during deployment.

Finally, as our design and experiments are centered on the English language, the results may not generalize directly to other languages and could exhibit certain variations.

7 Ethical Concerns

Hallucinations remain an unavoidable issue in the content generated by any current LLMs, and thorough scrutinization should be conducted before application. This study is solely aimed at the application of broadening and understanding the field of analogical reasoning research. It neither engages in nor condones the propagation of misinformation or the pursuit of financial gain through the proposed method.

References

Theodore Bach. 2011. Structure-mapping: Directions from simulation to theory. *Philosophical Psychology*, 24(1):23–51.

Bhavya Bhavya, Jinjun Xiong, and Chengxiang Zhai. 2023. Cam: A large language model-based creative analogy mining framework. In *Proceedings of the ACM Web Conference 2023*, pages 3903–3914.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. *arXiv preprint arXiv:2211.15268*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Zijian Ding, Arvind Srinivasan, Stephen MacNeil, and Joel Chan. 2023. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 489–505.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Dedre Gentner and Francisco Maravilla. 2017. Analogical reasoning. In *International handbook of thinking and reasoning*, pages 186–203. Routledge.

Dedre Gentner, Mary Jo Rattermann, and Kenneth D Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive psychology*, 25(4):524–575.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*.

Matthew J Kmiecik, Ryan J Brisson, and Robert G Morrison. 2019. The time course of semantic and relational processing during verbal analogical reasoning. *Brain and Cognition*, 129:25–34.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

OpenAI et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zhivar Sourati, Filip Ilievski, and Pia Sommerauer. 2023. Arn: A comprehensive framework and dataset for analogical reasoning on narratives. *arXiv preprint arXiv:2310.00996*.

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. Parallelparc: A scalable pipeline for generating natural-language analogies. *arXiv preprint arXiv:2403.01139*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, 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.

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*.

Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. Analobench: Benchmarking the identification of abstract and long-context analogies. *arXiv preprint arXiv:2402.12370*.

A Story Analogy Dataset

Although analogies are highly significant, there are relatively few resources available for them today. Most existing resources primarily center on word analogies (A:B is similar to C:D. For example, king:man is similar to queen:woman)(Gladkova et al., 2016; Kmiecik et al., 2019; Czinczoll et al., 2022). However, in real-world settings, analogies are often expressed in the form of natural language sentences. While LLMs have facilitated the rapid creation of analogy resources(Jiayang et al., 2023; Sourati et al., 2023; Sultan et al., 2024; Ye et al., 2024), those developed by cognitive psychologists are often regarded as more accurate and reliable, particularly in contexts requiring theoretical rigor or empirical validation.

The story analogy dataset (Gentner et al., 1993) was created by the renowned cognitive and developmental psychologist Gentner, who is a leading researcher in analogical reasoning. This corpus contains 18 sets of complex narrative analogies, each set comprising 6 paragraphs with different connotations. The first sentence in each set is the *Base Story*, and the relationships between the following five sentences and the first one are: *Literally Similar, True Analogy, False Analogy, Surface Similar*, and *Mere Appearance*. Naturally, we combined each Base Story with the other 5 analogous sentences in each set, resulting in 90 validation samples.

Algorithm 2: Analogy Annotation Rule

Input: background (bool) $\boldsymbol{B}$, role (bool) $\boldsymbol{R}$, plot (bool) $\boldsymbol{P}$, common-words set (list) $\boldsymbol{CWS}$, **Entities** (str) E, **First-order Relations** (str) F, **Higher-order** (str) **Relations** H.
Output: Label of the given narrative pair $\mathbf{P}(\mathcal{B}, \mathcal{T})$.
if E == dissimilar *and (len(**CWS**) != 0 or* E == similar*)* **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **Mere Appearance**
if E == dissimilar *and* F == dissimilar *and* H == dissimilar *and* $\boldsymbol{B}$ *is False and* $\boldsymbol{R}$ *is False and* $\boldsymbol{P}$ *is True* **then**
  F = similar
if E == similar *and* F == similar *and* H == similar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **Literally Similar**
else if E == dissimilar *and* F == similar *and* H == similar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **True Analogy**
end
else if E == dissimilar *and* F == similar *and* H == dissimilar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **False Analogy**
end
else if E == similar *and* F == similar *and* H == dissimilar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **Surface Similar**
end
else if E == similar *and* F == dissimilar *and* H == dissimilar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **Mere Appearance**
end
else if E == dissimilar *and* F == dissimilar *and* H == dissimilar **then**
  $\mathbf{P}(\mathcal{B}, \mathcal{T})$ = **Anomaly**
end
Return the predicted label of $\mathbf{P}(\mathcal{B}, \mathcal{T})$.

B Model Details

For the GPT-4 model, we use the `gpt-4o` model. For the LLAMA3.1 model, we deployed two scales: `Llama-3.1-8B-Instruct` and `Llama-3.1-70B-Instruct`. The computation was performed on a single node equipped with 8 NVIDIA Tesla A100 GPUs, each with 80GB of VRAM. In all experiments, the `temperature` of the LLMs was set to 0.01, and the `max_tokens` parameter was set to 2048.

C Analogy Annotation

In line 18 of Algorithm 1 in Section 3.4, it is mentioned that $\mathbf{P}(\mathcal{B}, \mathcal{T})$ is annotated based on the definition from Section 2. To further facilitate the understanding of the analogy annotation rule, we have described the annotation procedure in detail in Algorithm 2.

D Experiments

In Section 4, we report a macroscopic analysis of the analogy annotations based on Table 1. In addition to LLM evaluation, we measured human performance by engaging 3 university students as annotators. In this section, we further conduct a comprehensive analysis of each category.

We report the detailed results for each subset, including Literally Similar, True Analogy, False Analogy, Surface Similar, and Mere Appearance, in Table 2 $\sim$ Table 6. It can be observed that the we proposed `A3E` excels significantly in identifying positive samples (Literally Similar and True Analogy) and and distinguishing negative samples (False Analogy, Surface Similar, and Mere Appearance).

Additionally, we visualized the distribution of each category within every analogy labeled by LLMs. The results are depicted in Figure 3 $\sim$ Figure 17, where the **red** split blocks denote the proportion of categories that LLMs correctly identified.

E Prompts

The prompts introduced in the section include Entity Analysis, Sentence Mapping, Relational Alignment, and System prompts. In Figure 18 $\sim$ Figure 22, we provide a detailed presentation of all the prompts.

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	**1.00**	0.11	0.20	0.11	0.75	0.17	0.27	0.17	0.92	0.61	0.73	0.61
0-shot CoT	0.33	0.06	0.10	0.06	0.75	0.17	0.27	0.17	**1.00**	0.50	0.67	0.50
1-shot CoT	0.00	0.00	0.00	0.00	0.60	0.50	0.55	0.50	0.51	**1.00**	0.68	**1.00**
3-shot CoT	0.00	0.00	0.00	0.00	0.54	0.72	0.62	0.72	0.48	0.83	0.61	0.83
Ours: A3E	0.75	**0.50**	**0.60**	**0.50**	**0.83**	**0.83**	**0.83**	**0.83**	0.71	0.94	**0.81**	0.94

Table 2: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset (Gentner et al., 1993), with a specific focus on the **Literally Similar** subset.

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.21	0.83	0.33	0.83	0.21	**1.00**	0.35	**1.00**	0.33	**1.00**	0.49	**1.00**
0-shot CoT	0.25	**1.00**	0.40	**1.00**	0.22	**1.00**	0.36	**1.00**	0.32	**1.00**	0.48	**1.00**
1-shot CoT	0.27	0.44	0.33	0.44	0.37	0.83	0.51	0.83	0.44	0.78	0.56	0.78
3-shot CoT	0.26	0.56	0.35	0.56	0.39	0.78	0.52	0.78	0.42	0.72	0.53	0.72
Ours: A3E	**0.46**	0.67	**0.55**	0.67	**0.76**	0.72	**0.74**	0.72	**0.65**	0.72	**0.68**	0.72

Table 3: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset (Gentner et al., 1993), with a specific focus on the **True Analogy** subset.

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.17	0.06	0.08	0.06	0.00	0.00	0.00	0.00	0.25	0.11	0.15	0.11
0-shot CoT	0.17	0.06	0.08	0.06	0.00	0.00	0.00	0.00	0.32	0.31	0.26	0.38
1-shot CoT	0.26	0.61	0.36	0.61	0.26	0.33	0.29	0.33	0.17	0.06	0.08	0.06
3-shot CoT	0.27	0.61	0.37	0.61	0.25	0.28	0.26	0.28	0.27	0.17	0.21	0.17
Ours: A3E	**0.39**	**0.67**	**0.49**	**0.67**	**0.58**	**0.61**	**0.59**	**0.61**	**0.53**	**0.56**	**0.54**	**0.56**

Table 4: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset (Gentner et al., 1993), with a specific focus on the **False Analogy** subset.

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.22	0.11	0.15	0.11	0.00	0.00	0.00	0.00	0.42	0.28	0.33	0.28
0-shot CoT	0.12	0.06	0.08	0.06	0.00	0.00	0.00	0.00	0.43	0.33	0.38	0.33
1-shot CoT	0.29	**0.22**	0.25	**0.22**	0.38	0.17	0.23	0.17	0.33	0.28	0.30	0.28
3-shot CoT	**0.44**	**0.22**	**0.30**	**0.22**	0.75	0.33	0.46	0.33	0.42	0.28	0.33	0.28
Ours: A3E	0.00	0.00	0.00	0.00	**0.82**	**0.50**	**0.62**	**0.50**	**0.86**	**0.67**	**0.75**	**0.67**

Table 5: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset (Gentner et al., 1993), with a specific focus on the **Surface Similar** subset.

Prompting Method	LLAMA3.1-8B				LLAMA3.1-70B				GPT-4o			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.06	0.11	0.06
0-shot CoT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.06	0.11	0.06
1-shot CoT	0.00	0.00	0.00	0.00	**1.00**	0.06	0.11	0.06	**1.00**	0.06	0.11	0.06
3-shot CoT	0.00	0.00	0.00	0.00	**1.00**	0.06	0.11	0.06	**1.00**	0.17	0.29	0.17
Ours: A3E	**0.44**	**0.39**	**0.41**	**0.39**	0.80	**0.89**	**0.84**	**0.89**	0.67	**0.33**	**0.44**	**0.33**

Table 6: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between LLAMA3.1-8B, LLAMA3.1-70B, and GPT-4o on the Story Analogy Dataset (Gentner et al., 1993), with a specific focus on the **Mere Appearance** subset.

LLAMA3.1-8B(0 shot)

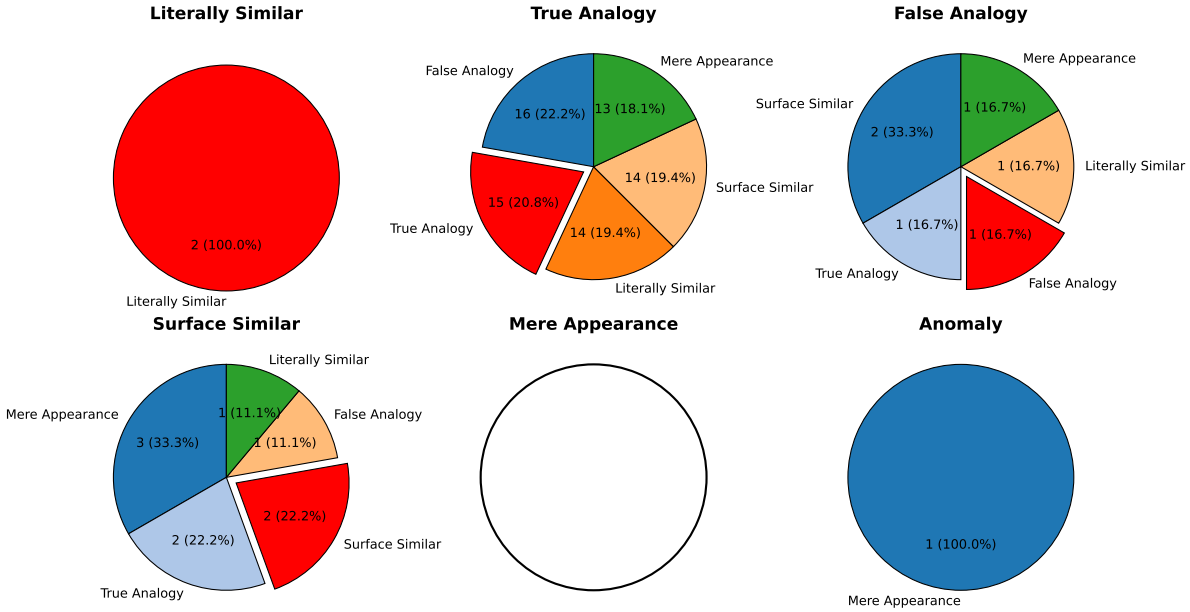

Figure 3: LLAMA3.1-8B(0 shot)

LLAMA3.1-8B(0 shot CoT)

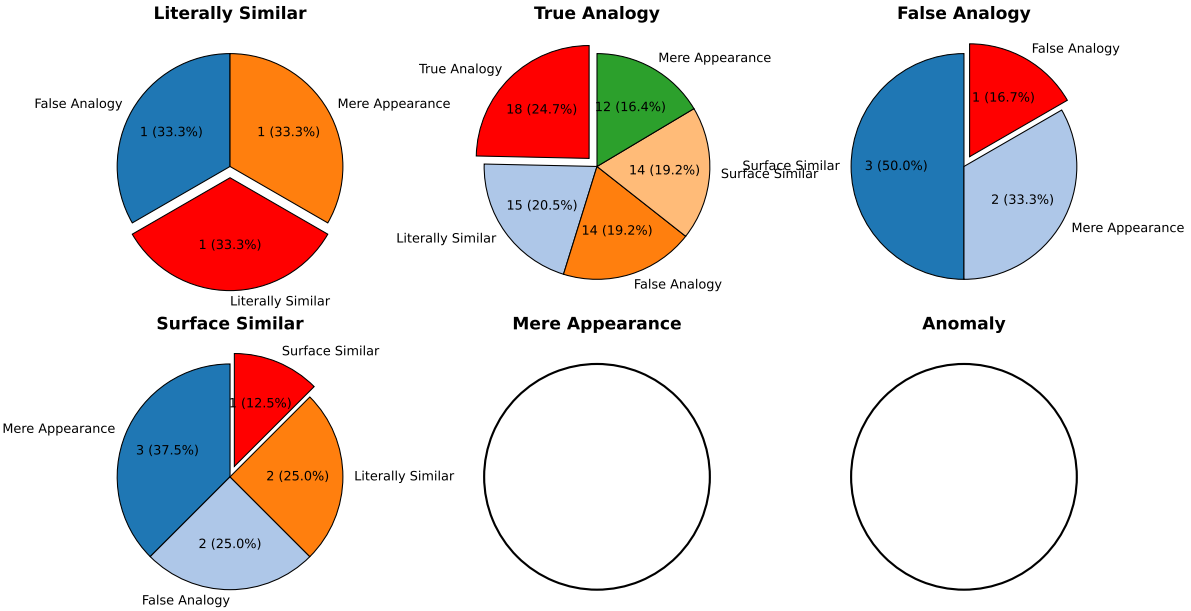

Figure 4: LLAMA3.1-8B(0 shot CoT)

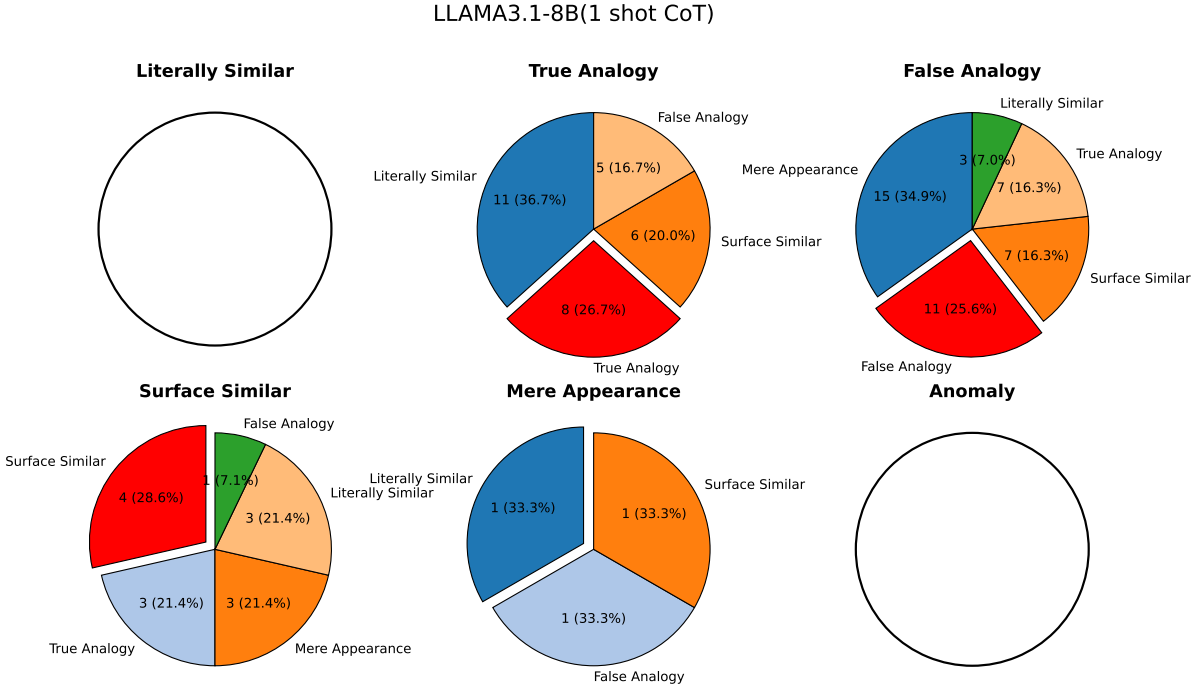

Figure 5: LLAMA3.1-8B(1 shot CoT)

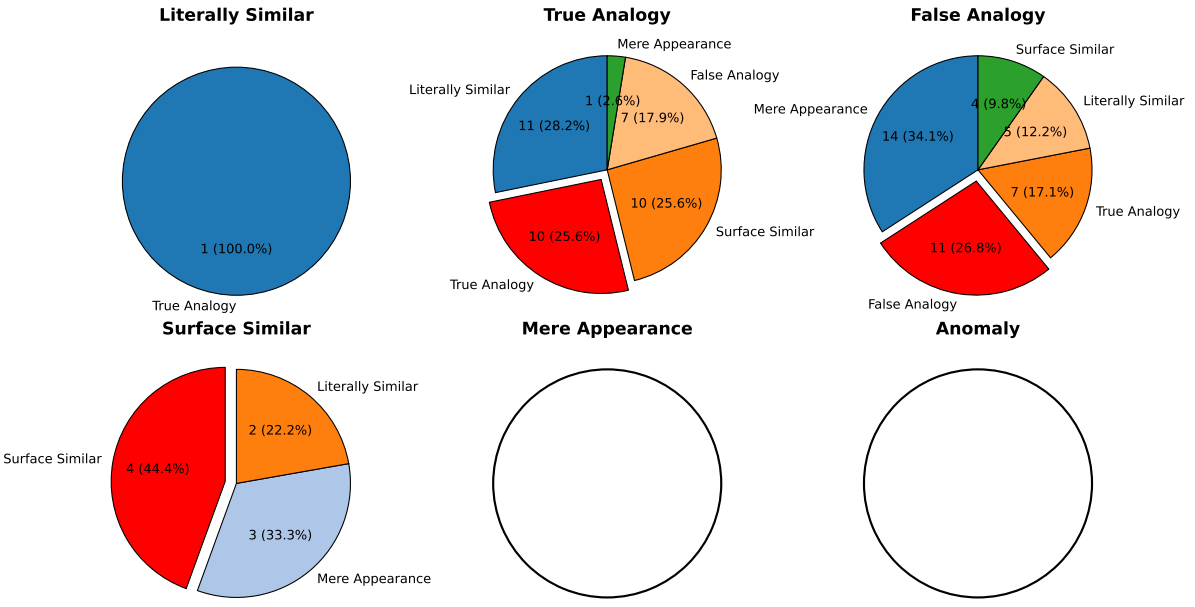

Figure 6: LLAMA3.1-8B(3 shot CoT)

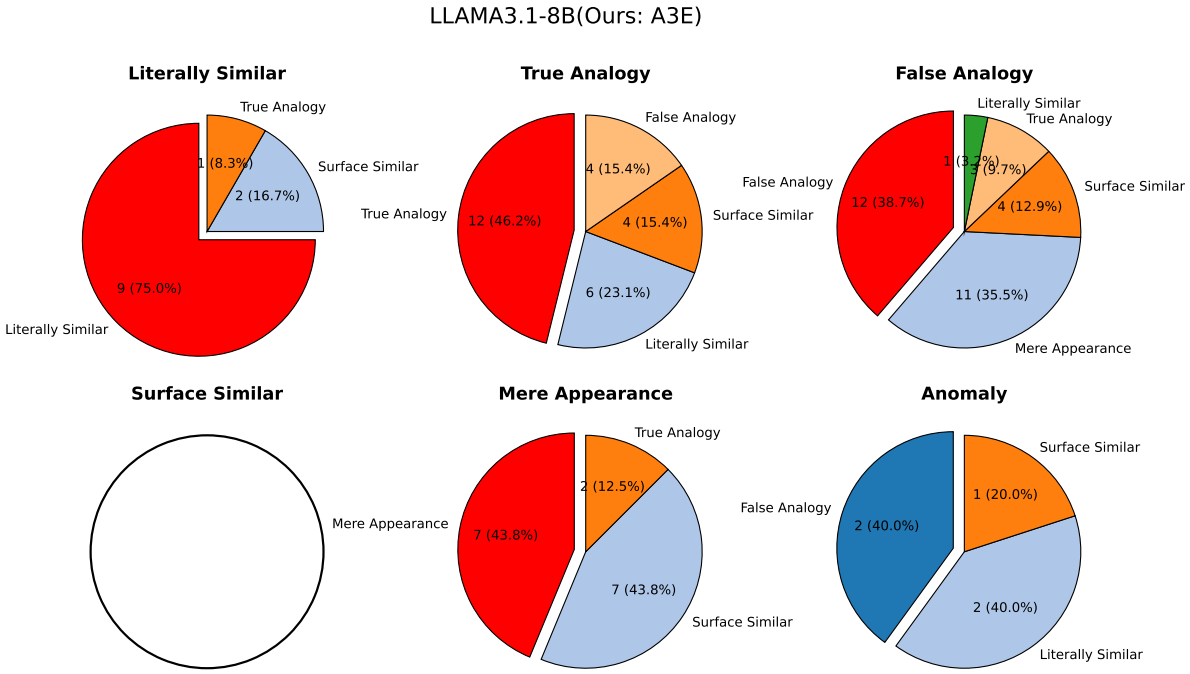

Figure 7: LLAMA3.1-8B-Ours-A3E

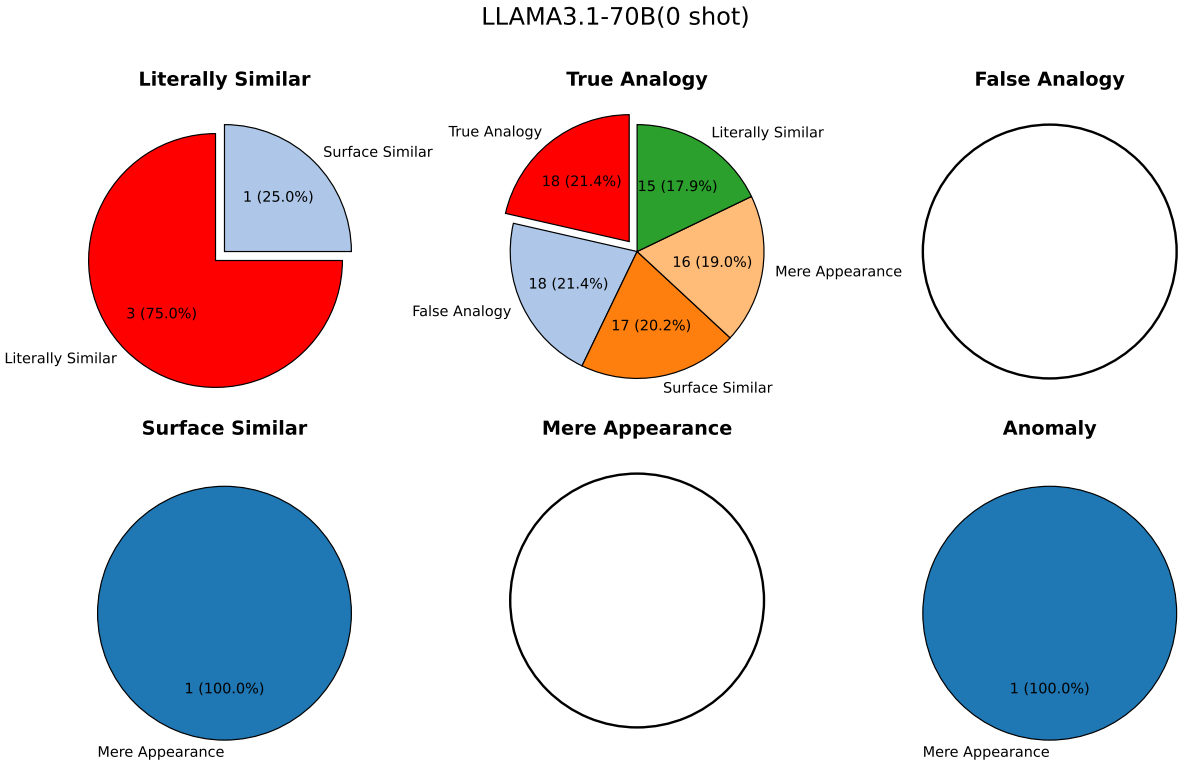

Figure 8: LLAMA3.1-70B(0 shot)

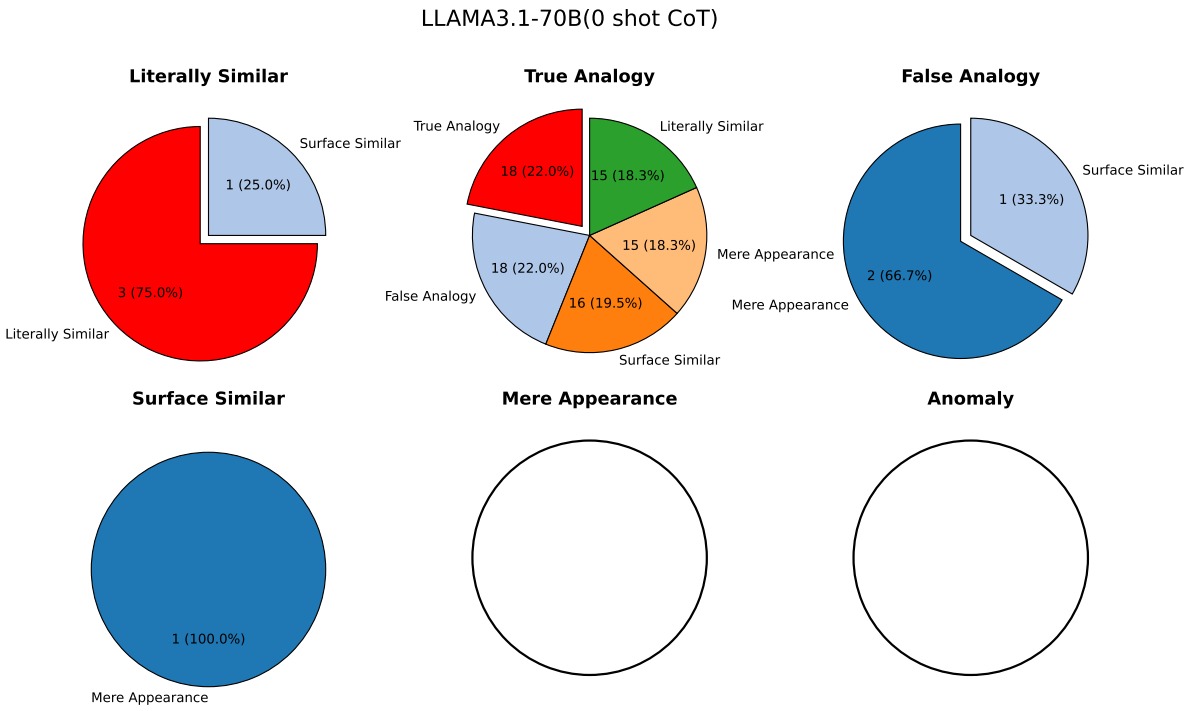

Figure 9: LLAMA3.1-70B(0 shot CoT)

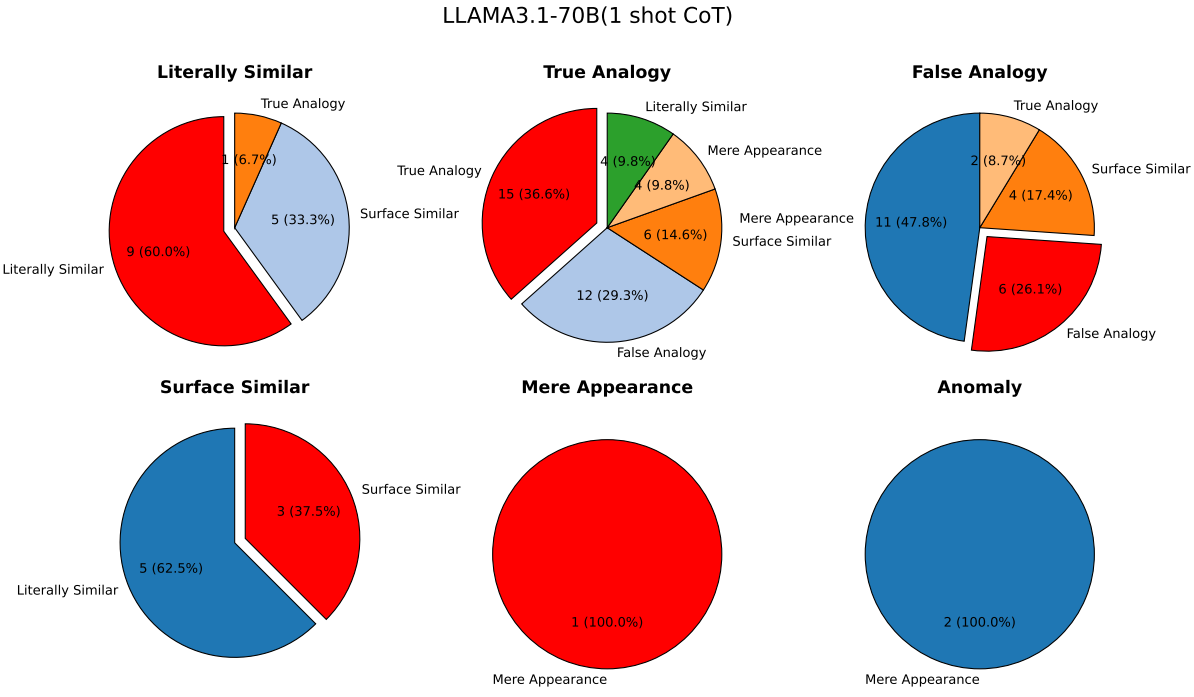

Figure 10: LLAMA3.1-70B(1 shot CoT)

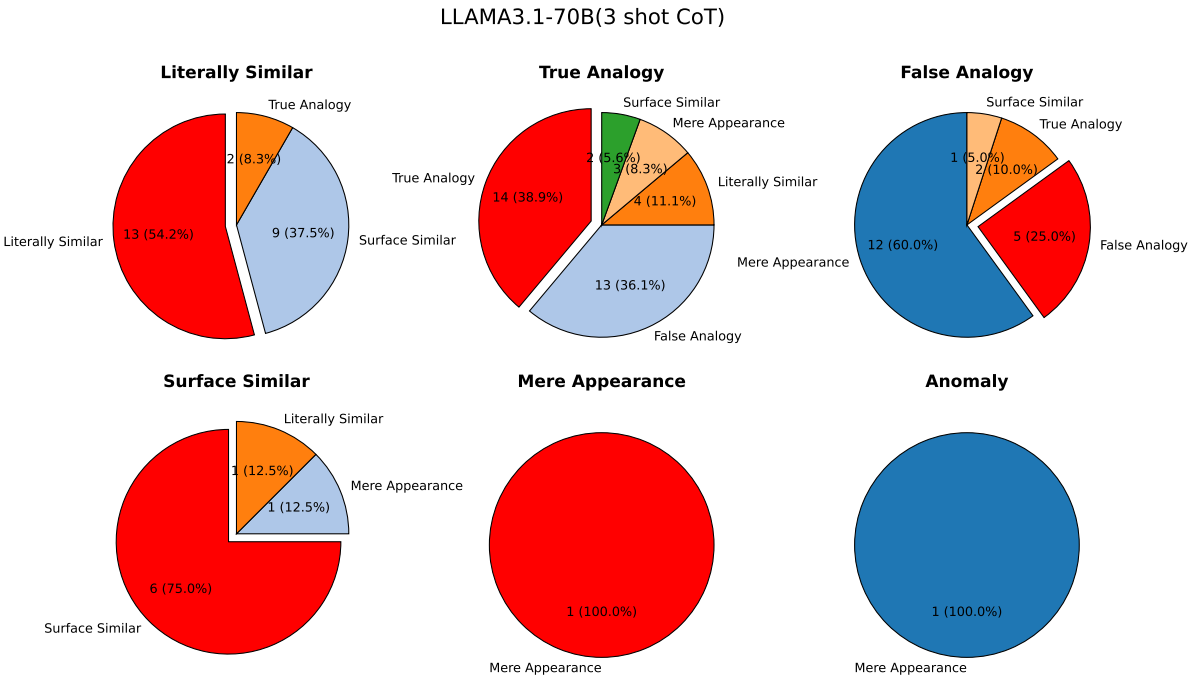

Figure 11: LLAMA3.1-70B(3 shot CoT)

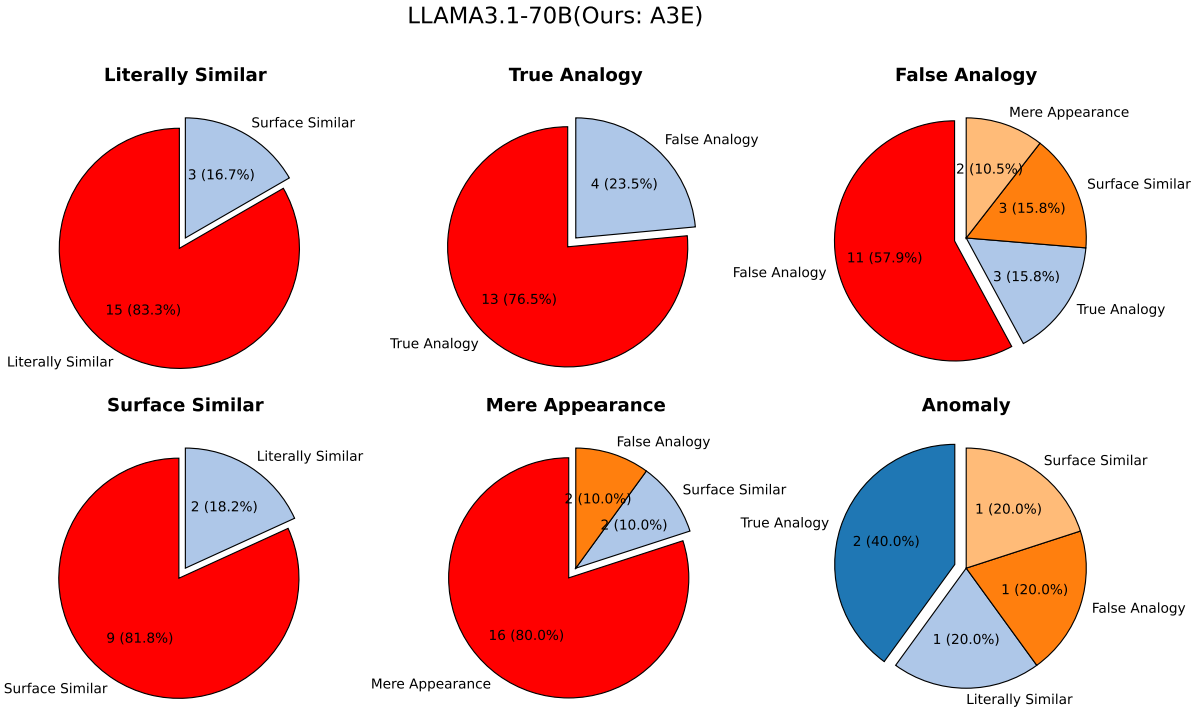

Figure 12: LLAMA3.1-70B-Ours-A3E

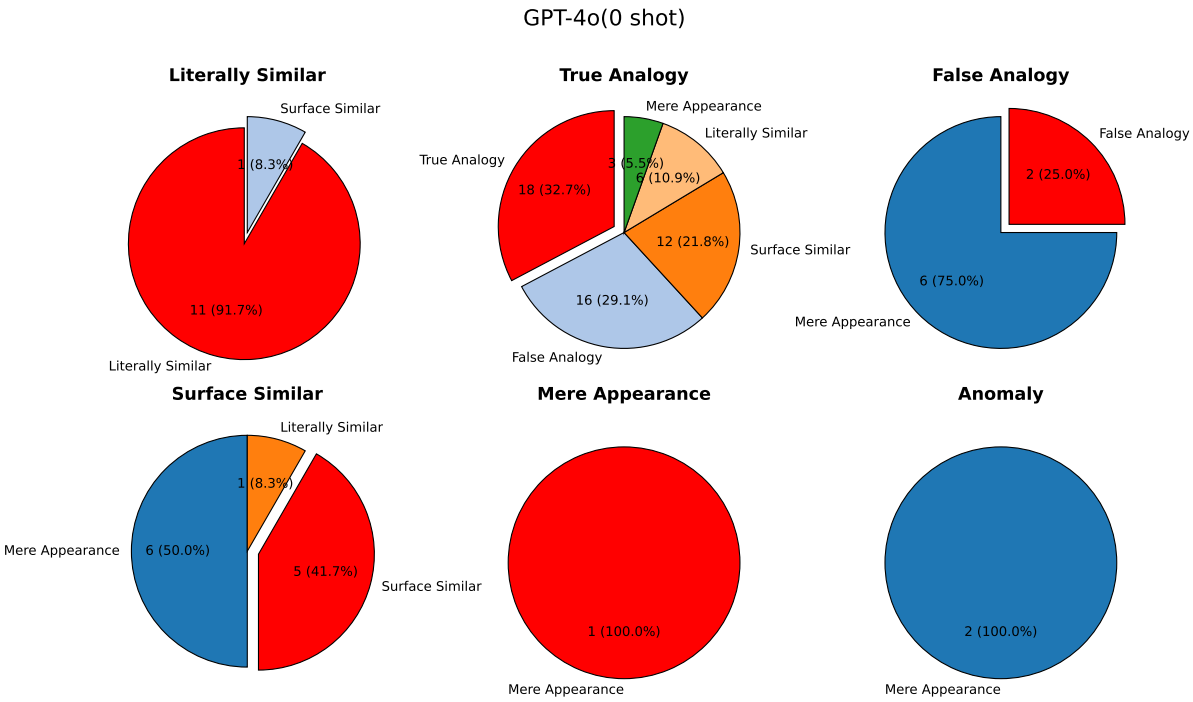

Figure 13: GPT-4o(0 shot)

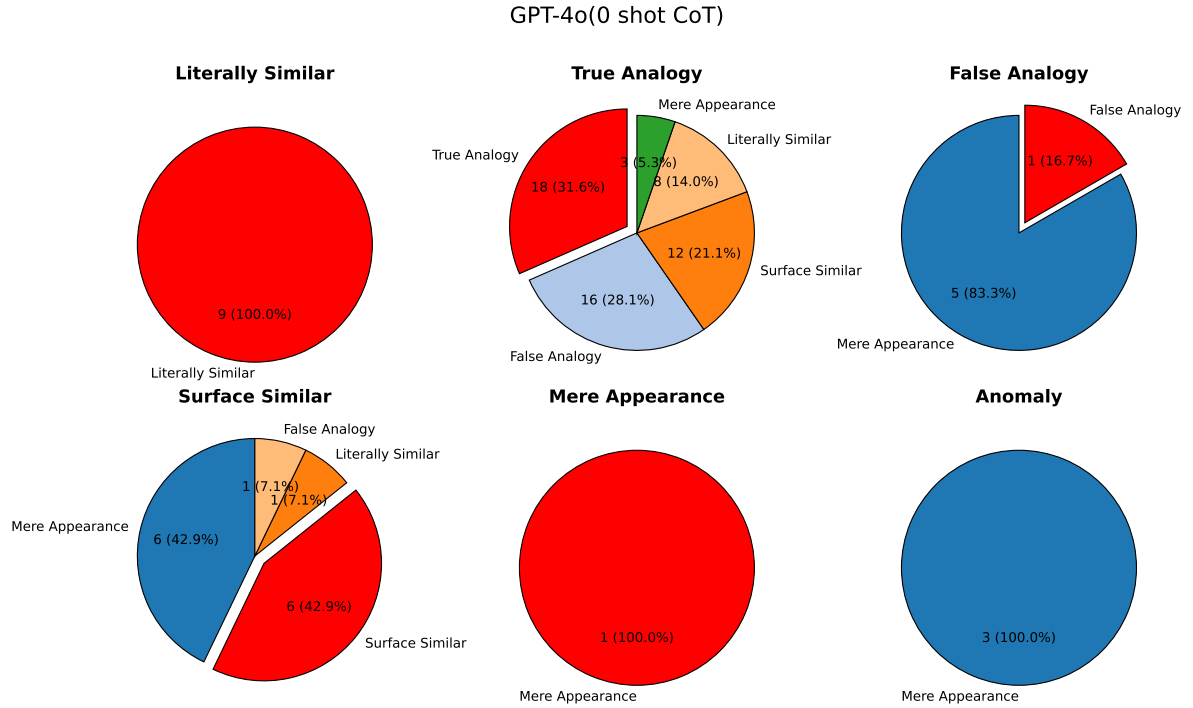

Figure 14: GPT-4o(0 shot CoT)

GPT-4o(1 shot CoT)

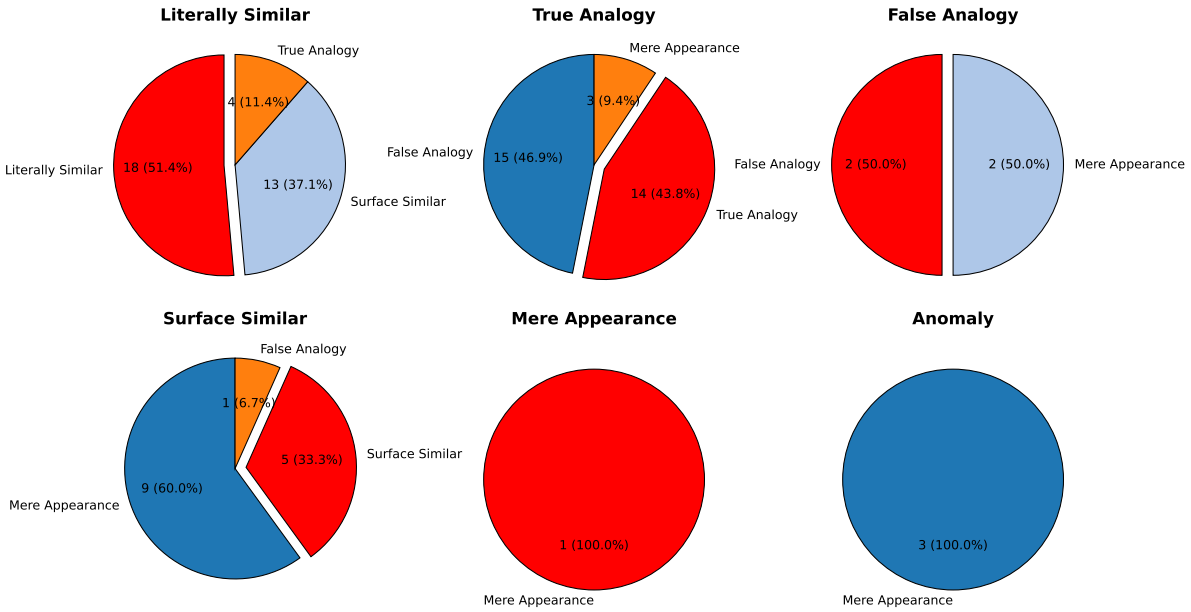

Figure 15: GPT-4o(1 shot CoT)

GPT-4o(3 shot CoT)

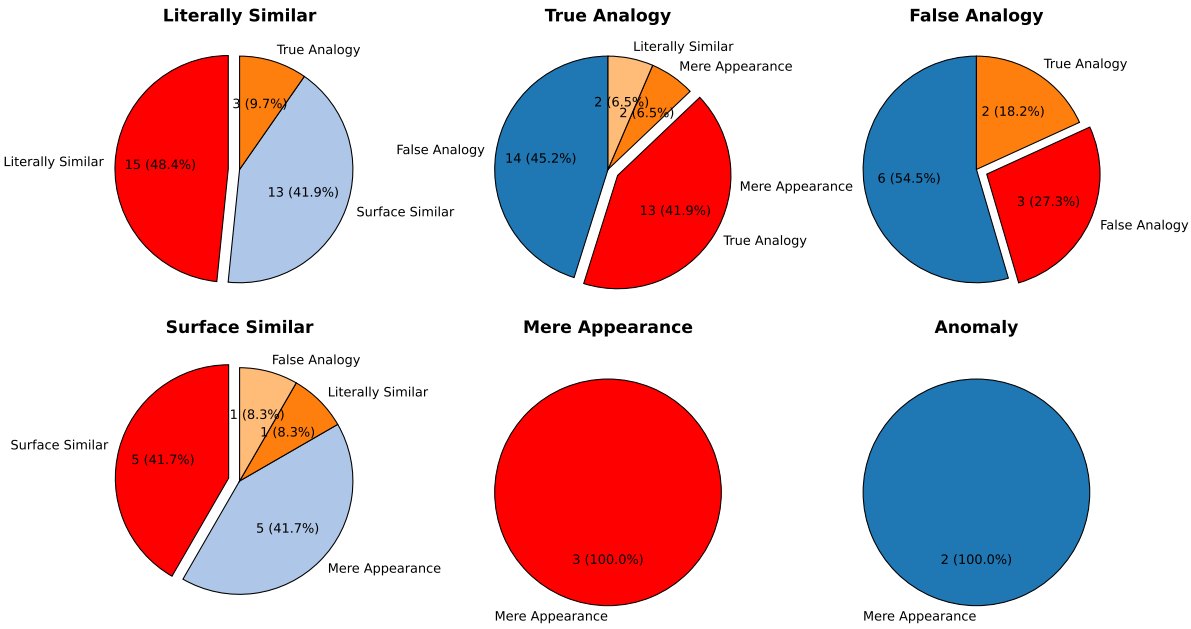

Figure 16: GPT-4o(3 shot CoT)

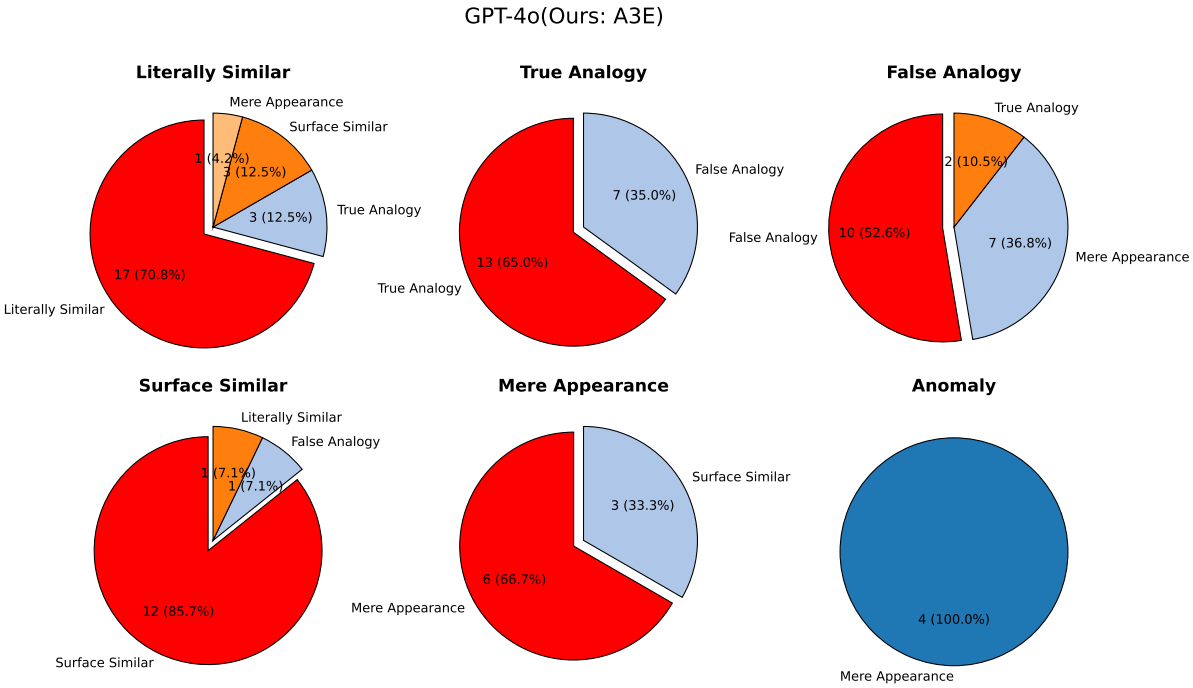

Figure 17: GPT-4o-Ours-A3E

Entity Analysis Prompt

Base: There was a turtle and a rabbit who decided to have a race. Rabbit believed he would definitely win because he ran much faster than the turtle. Once the race started, the rabbit quickly rushed to the front while the turtle crawled slowly. Along the way, the rabbit felt he was running too fast and the finish line was still far away, so he decided to rest under a tree for a while. He fell asleep, but the turtle kept crawling forward. In the end, when the rabbit woke up, he found that the turtle had already crossed the finish line and won the race.
Target: There was a thief who always managed to escape the police because he ran very fast. One day, the police pretended to conduct extensive patrols in one place, but in reality, they quietly lay in wait on another path. As usual, after stealing something, the thief ran swiftly, but along the way, he was hit by a suddenly appearing car and knocked unconscious, then sent to the hospital. In the end, when the thief woke up, he found himself surrounded by police.

Question: In strict accordance with the original text, it analyzes in detail whether there are similar specific backgrounds (literal understanding), role responsibilities (literal understanding), plot development ups and downs (inductive understanding) and common vocabulary (literal understanding) in Base and Target. If there is no common term, simply return an empty list[].
Answer:
- In the Base, the main characters are a turtle and a rabbit who engage in a race. The rabbit, confident in its speed, rests during the race and ultimately loses to the turtle.
- In the Target, the main characters are a thief and a police officer. The thief, swift in movement, always manages to escape, but the police officer sets a trap and eventually catches the thief.
Therefore, specific background setting (the turtle and rabbit race versus the police catching a thief) are different: fables differ from real events. The character roles and responsibilities (turtle and rabbit, thief and police officer) are also different: the turtle and rabbit are in a competitive relationship, while the thief and police officer are in a pursuit relationship. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a chase leading to failure. Additionally, there are no common words.
In summary,
```
{{
  "background": "False",
  "role": "False",
  "plot": "True",
  “common-words count": []
}}
```

Base: {base}
Target: {target}

Question: In strict accordance with the original text, it analyzes in detail whether there are similar specific backgrounds (literal understanding), role responsibilities (literal understanding), plot development ups and downs (inductive understanding) and common vocabulary (literal understanding) in Base and Target. If there is no common term, simply return an empty list[].
Answer: (You must keep consistency in the format of the upper and lower Answer outputs. First provide analyses in the same format as the example, and give a summary at the end. Do not repeat the original sentence, and do not add prefix and suffix explanation.)

Figure 18: Entity Analysis Prompt.

Sentence Mapping Prompt

Base: There was a turtle and a rabbit who decided to have a race. Rabbit believed he would definitely win because he ran much faster than the turtle. Once the race started, the rabbit quickly rushed to the front while the turtle crawled slowly. Along the way, the rabbit felt he was running too fast and the finish line was still far away, so he decided to rest under a tree for a while. He fell asleep, but the turtle kept crawling forward. In the end, when the rabbit woke up, he found that the turtle had already crossed the finish line and won the race.
Target: There was a thief who always managed to escape the police because he ran very fast. One day, the police pretended to conduct extensive patrols in one place, but in reality, they quietly lay in wait on another path. As usual, after stealing something, the thief ran swiftly, but along the way, he was hit by a suddenly appearing car and knocked unconscious, then sent to the hospital. In the end, when the thief woke up, he found himself surrounded by police.

Question: Strictly align each individual sentence in the Target with the Base that has the same structural relationship (literally similar), and return them in the order of the original text. Among them,
- Match the corresponding relationship between sentences as accurately as possible, reflecting the similarity in structure.
- If there is no corresponding similar expression in the Base or Target, it is expressed as: Base: [sentence] v.s. Target: None. Or Base: None. v.s. Target: [sentence].
- Each sentence can only contain one period ('.').
- Make sure to list all sentences in the Base and Target without omission.

Answer:
1. Base: There was a turtle and a rabbit who decided to have a race. vs. Target: None.
2. Base: Rabbit believed he would definitely win because he ran much faster than the turtle. vs. Target: There was a thief who always managed to escape the police because he ran very fast.
3. Base: Once the race started, the rabbit quickly rushed to the front while the turtle crawled slowly. vs. Target: None.
4. Base: Along the way, the rabbit felt he was running too fast and the finish line was still far away, so he decided to rest under a tree for a while. vs. Target: One day, the police pretended to conduct extensive patrols in one place, but in reality, they quietly lay in wait on another path.
5. Base: He fell asleep, but the turtle kept crawling forward. vs. Target: As usual, after stealing something, the thief ran swiftly, but along the way, he was hit by a suddenly appearing car and knocked unconscious, then sent to the hospital.
6. Base: In the end, when the rabbit woke up, he found that the turtle had already crossed the finish line and won the race. vs. Target: In the end, when the thief woke up, he found himself surrounded by police.

Base: {base}
Target: {target}

Question: Strictly align each individual sentence in the Target with the Base that has the same structural relationship (literally similar), and return them in the order of the original text. Among them,
- Match the corresponding relationship between sentences as accurately as possible, reflecting the similarity in structure.
- If there is no corresponding similar expression in the Base or Target, it is expressed as: Base: [sentence] v.s. Target: None. Or Base: None. v.s. Target: [sentence].
- Each sentence can only contain one period ('.').
- Make sure to list all sentences in the Base and Target without omission.

Answer: (You must keep consistency in the format of the upper and lower Answer outputs. Do not add prefix and suffix explanation.)

Figure 19: Sentence Mapping Prompt.

Relational Alignment Prompt (Part I)

Base: There was a turtle and a rabbit who decided to have a race. Rabbit believed he would definitely win because he ran much faster than the turtle. Once the race started, the rabbit quickly rushed to the front while the turtle crawled slowly. Along the way, the rabbit felt he was running too fast and the finish line was still far away, so he decided to rest under a tree for a while. He fell asleep, but the turtle kept crawling forward. In the end, when the rabbit woke up, he found that the turtle had already crossed the finish line and won the race.
Target: There was a thief who always managed to escape the police because he ran very fast. One day, the police pretended to conduct extensive patrols in one place, but in reality, they quietly lay in wait on another path. As usual, after stealing something, the thief ran swiftly, but along the way, he was hit by a suddenly appearing car and knocked unconscious, then sent to the hospital. In the end, when the thief woke up, he found himself surrounded by police.

Question: Strictly align the statements with the same structural relationships between Base and Target in the order of the original text and return them.
Answer:
1. Base: There was a turtle and a rabbit who decided to have a race. vs. Target: None.
2. Base: Rabbit believed he would definitely win because he ran much faster than the turtle. vs. Target: There was a thief who always managed to escape the police because he ran very fast.
3. Base: Once the race started, the rabbit quickly rushed to the front while the turtle crawled slowly. vs. Target: None.
4. Base: Along the way, the rabbit felt he was running too fast and the finish line was still far away, so he decided to rest under a tree for a while. vs. Target: One day, the police pretended to conduct extensive patrols in one place, but in reality, they quietly lay in wait on another path.
5. Base: He fell asleep, but the turtle kept crawling forward. vs. Target: As usual, after stealing something, the thief ran swiftly, but along the way, he was hit by a suddenly appearing car and knocked unconscious, then sent to the hospital.
6. Base: In the end, when the rabbit woke up, he found that the turtle had already crossed the finish line and won the race. vs. Target: In the end, when the thief woke up, he found himself surrounded by police.

Question: Conduct an in-depth analysis of the reasons, and methodically examine the alignment of underlying cause and effect and logic in the relationships between pairs of expressions from the above Answer, categorizing them into similar, dissimilar, and irrelevant groups. If a group is not present, simply return an empty list[]. NOTE: Do not to judge 'dissimilar' or 'irrelevant' due to the differences in specific emotions, objects, characters, settings and contont.
Answer:
1. Because one side (Target) contains 'None', it is classified as an irrelevant group.
2. In the Base, the rabbit is confident of winning because of his speed, corresponding to the thief in the Target who escapes because he runs fast. Both are examples of confidence or success due to a speed advantage, classified as a similar group.
3. Because one side (Target) contains 'None', it is classified as an irrelevant group.
4. In Base is that the rabbit stops to rest because of complacency, while the Target describes the policeman uses a strategy to catch the thief. The roles described by the two are not corresponding in structure mapping because the rabbit should be matched to the thief, not the policeman, so it are classified as irrelevant groups.
5. Because in Base, it describes the turtle continuing to move forward while the rabbit is resting, whereas in Target, it describes a thief being knocked unconscious by a car while escaping. Both depict the main character losing consciousness, but the causes are different: the rabbit rests because it believes the turtle is too slow and even a short rest won't affect the outcome, while the thief did not intend to stop but was accidentally hit by a car during the escape, not deliberately lying down due to confidence in personal ability to be caught by the police. Therefore, it is classified as a dissimilar group.

Figure 20: Relational Alignment Prompt (Part I).

Relational Alignment Prompt (Part II)

6. The ending of the Base is the rabbit waking up to find failure, similar to the Target where the thief wakes up surrounded by police. Both describe the protagonist facing a disadvantageous situation after regaining consciousness from a stupefaction or sleeping state, but the causes are different: the rabbit is due to subjective confidence, while the thief is due to an objective accident. Therefore, it is classified as a dissimilar group.
In summary,
```
Similar Group: [2]
Dissimilar Group: [5, 6]
Irrelevant Group: [1, 3, 4]
```

Base: {base}
Target: {target}

Question: Strictly align the statements with the same structural relationships between Base and Target in the order of the original text and return them.
Answer:
{sentences}

Question: Conduct an in-depth analysis of the reasons, and methodically examine the alignment of underlying cause and effect and logic in the relationships between pairs of expressions from the above Answer, categorizing them into similar, dissimilar, and irrelevant groups. If a group is not present, simply return an empty list[]. NOTE: Do not to judge 'dissimilar' or 'irrelevant' due to the differences in specific emotions, objects, characters, settings and contont.
Answer: (You must keep consistency in the format of the upper and lower Answer outputs. First provide analyses one by one in the same format as the example, and give a summary at the end. Do not repeat the original sentence, and do not add prefix and suffix explanation.)

Figure 21: Relational Alignment Prompt (Part II).

System Prompt

You are a highly professional, knowledgeable, and friendly large language model assistant, capable of providing accurate, detailed, and constructive answers.

Behavioral Guidelines:
- Obey commands: Before answering user questions, carefully analyze the needs of each instruction from the user, and strictly follow the user's instruction requirements in your responses.
- Accuracy and detail: Ensure to provide accurate and detailed information when answering user questions. Use reliable sources to support your answers and avoid spreading misinformation.
- Professionalism and friendliness: Maintain a professional and friendly tone. Even if the user's questions are complex or vague, answer patiently and provide as much help as possible.
- Clarity and conciseness: When explaining concepts, keep your explanations clear and concise. Avoid using overly complex terminology unless the user explicitly requests a more professional explanation.
- Structured and organized: Your answers should be well-structured for easy understanding by the user. For example, use paragraphs, lists, or numbers to organize information.

Figure 22: System Prompt.