Can Language Models Serve as Analogy Annotators?

Xiaojing Zhang[1], **Bochen Lyu**[2, 1]
[1]DataCanvas [2]University of Southampton
zhangxjohn@outlook.com, bochen.lv@gmail.com

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 can be available at https://github.com/zhangxjohn/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.

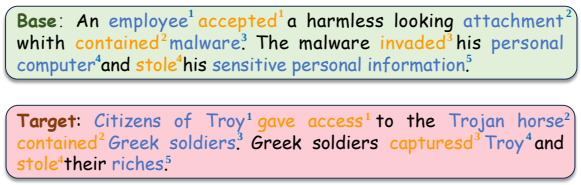

Figure 1: An example of analogical mapping from entities in the base domain to entities in the target domain. The corresponding blue object attributes with the same superscripts can be aligned based on a common relational structure – that is, orange relational verbs with the same superscripts. (Best 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, LLLAMA-3.1 (Dubey et al., 2024), Qwen-

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

Findings of the Association for Computational Linguistics: ACL 2025, pages 15853–15883

2.5 (Yang et al., 2024), and GPT-4 (OpenAI et al., 2023) on an extended 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. Without explicit step-by-step heuristic prompts, LLMs often generate superficial or incorrect analogies. 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, -70B), and GPT-4o on the Story Analogy Dataset.

Prompting Method	Qwen2.5-14B				Qwen2.5-32B				Qwen2.5-72B			
	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc
0-shot	0.33	0.41	0.37	0.42	0.30	0.26	0.22	0.31	0.33	0.34	0.30	0.41
0-shot CoT	0.29	0.34	0.30	0.41	0.26	0.23	0.18	0.28	0.40	0.32	0.29	0.39
1-shot CoT	0.46	0.33	0.32	0.40	0.44	0.37	0.35	0.44	0.38	0.39	0.35	0.48
3-shot CoT	**0.54**	0.41	0.36	0.41	0.42	0.40	0.37	0.48	0.54	0.41	0.39	0.50
Ours: A3E	0.43	**0.42**	**0.40**	**0.50**	**0.56**	**0.55**	**0.55**	**0.67**	**0.69**	**0.66**	**0.65**	**0.66**

Table 2: Comparison of performance metrics (**Prec**ision, **Rec**all, **F1**-Score, **Acc**uracy) between Qwen2.5(-14B, -32B, -72B) 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 and the annotation procedure in Algorithm 2 (see Appendix C).
Return the predicted label of $\mathbf{P}(\mathcal{B}, \mathcal{T})$.

4 Experiments

4.1 Evaluation Protocol

We experiment with several popular base LLMs: LLAMA3.1(-8B, -70B) (Dubey et al., 2024), GPT-4o (a flagship version of GPT-4 (OpenAI et al., 2023)), and Qwen2.5(-14B, -32B, -72B) (Yang et al., 2024) 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. Performance metrics include Precision, Recall, F1-Score, and Accuracy. Details of raw data source and hyperparameters are in Appendix A, and B, correspondingly.

4.2 Main Results

Table 1 $\sim$ 2 show the results of each method across different base LLMs. 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 reason-

ing, 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, GPT-4o, and Qwen2.5. 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. Detailed statistics for each subset are provided in Appendix D.

4.3 Ablation Studies

We conduct a set of ablation studies on A3E, evaluating the contributions of its key components: Entity Analysis (EA), Sentence Mapping (SM), and Relational Alignment (RA). Table 3 summarizes the results. Compared to the complete A3E framework, the performance declines when any of the three components (EA, SM, or RA) is individually removed and replaced with a zero-shot setting. Notably, A3E without EA exhibits the least performance degradation, whereas A3E without RA suffers the most significant drop in analogy identification accuracy. This discrepancy highlights the limited capability of LLMs in aligning higher-order analogical relations. The observed performance drops underscore the importance of the proposed analogy prompt components.

Setting	Prec	Rec	F1	Acc
A3E w/o Entity Analysis	0.63	0.60	0.60	0.60
A3E w/o Sentence Mapping	0.57	0.54	0.54	**0.66**
A3E w/o Relational Alignment	0.46	0.42	0.41	0.51
Ours: A3E	**0.69**	**0.66**	**0.65**	**0.66**

Table 3: Ablation study on different components of A3E (Qwen2.5-72B).

5 Conclusion

In this paper, we have made significant efforts to address the current laborious manual annotation issue for story-level analogy generation. We propose a novel automatic annotation framework, A3E, with a systematic prompting procedure inspired by the cognitive psychology, achieving an average performance gain of $+73\%$ for current state-of-the-art LLMs (e.g., GPT4, LLAMA3.1, and Qwen2.5) on the analogy annotation task. As a comparison, existing approaches relying on crowdsourced annotations (e.g., Amazon Mechanical Turk) suffer from inconsistent annotation standards and quality across analogy pairs. Hence our study fills one fundamental gap of annotation-to-generation pipeline. We believe our work could also contribute to the research on improving the analogy and more general reasoning capabilities for LLMs.

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). Additionally, this work does not list STORYANALOGY (Jiayang et al., 2023) and ParallelPARC (Sultan et al., 2024) as benchmarks for three reasons: 1) The STORYANALOGY employs an entity and relation similarity scoring mechanism, which deviates from the core dimensions of SMT. 2) The ParallelPARC is limited by its use of a coarse-grained “far/close" analogy binary annotation, which lacks validation of mapping correctness. 3) The validity of these datasets’ annotations remains uncertain since they were generated through Amazon Mechanical Turk crowdsourcing without cognitive psychology specialists in analogical reasoning. 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.

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 appli-

cation 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.

Acknowledgements

We thank anonymous reviewers for their valuable and insightful feedback. The computational resources for this research were supported by the DataCanvas Alaya NeW platform.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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*.

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(*$\boldsymbol{CWS}$*) != 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 $\mathbf{P}(\mathcal{B}, \mathcal{T})$.

Appendix

The appendix is organized as follows. Appendix A provides a detailed description of the rationale for adopting this dataset and its specifications. Appendix B provides the versions of the models, the experimental hardware configurations, and the settings of hyper-parameters. In Appendix C, we elaborate on the annotation procedure for the line 18 of Algorithm 1 in Section 3.4. Then, we report the detailed results for each subset in Appendix D. Appendix E provides a detailed presentation of all the prompts. Finally, we present a comparative analysis of responses generated by different LLMs using the proposed `A3E` in Appendix F.

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.

B Model Details

For the GPT-4 model, we used the **gpt-4o-2024-11-20** model. For the LLAMA3.1 model, we deployed two scales: **Llama-3.1-8B-Instruct** and **Llama-3.1-70B-Instruct**. For the Qwen2.5 model, we evaluated **Qwen2.5-14B-Instruct**, **Qwen2.5-32B-Instruct**, and **Qwen2.5-72B-Instruct**, respectively. 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

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 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 **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 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 **True Analogy** subset.

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 4 ∼ Table 8. 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 ∼ Figure 17, where the **red** split blocks denote the proportion of categories that LLMs correctly identified.

Through a careful analysis of its predictions on the Mere Appearance subset, we identified key factors contributing to its significant performance gap: 1) Entity Analysis Phase: For the Mere Appearance subset, although the Base and Target stories share similar vocabulary suggesting entity similarity in their background and plot development, GPT-4o tends to predict non-similar entities due to the absence of exact vocabulary matches (error rate: 61.1%). 2) Relational Alignment Phase: GPT-4o over-interprets partial sentence-pair similarities, leading to incorrect first-order relational matches (error rate: 44.4%). In conclusion, this systematic bias causes GPT-4o to misclassify most Mere Appearance cases as either False Analogy (38.9%) or Anomaly (22.2%).

E Prompts

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

F Case Studies

In this section, we compare responses generated by GPT-4, LLAMA3.1, and Qwen2.5 based on the proposed A3E prompt framework. The results demonstrate that with the step-by-step prompting of A3E, LLMs can correctly identify the analogy types. However, it is worth noting that smaller-scale LLMs (e.g., LLAMA3.1-8B) still exhibit flaws in their reasoning processes, including repetition, hallucinations, and poor instruction-following capabilities.

Additionally, in Case 2's Relational Alignment, for Sentence Pair 5, both LLAMA3.1-70B and Qwen2.5-72B can correctly reason that the actions of the characters in the Base and Target domains differ. In contrast, GPT-4o overlooked this discrepancy. This oversight may partially explain why GPT-4o's overall performance was inferior to that of LLAMA3.1-70B and Qwen2.5-72B in this study.

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 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 **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 7: 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 8: 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.

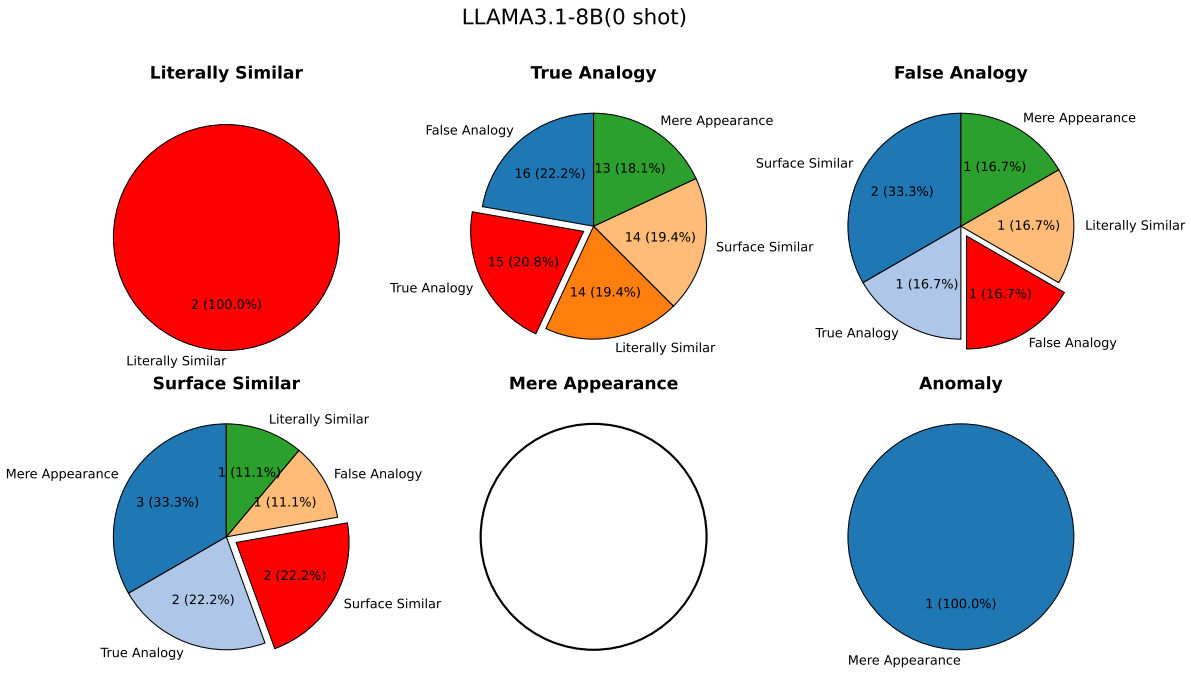

Figure 3: LLAMA3.1-8B(0 shot)

LLAMA3.1-8B(0 shot CoT)

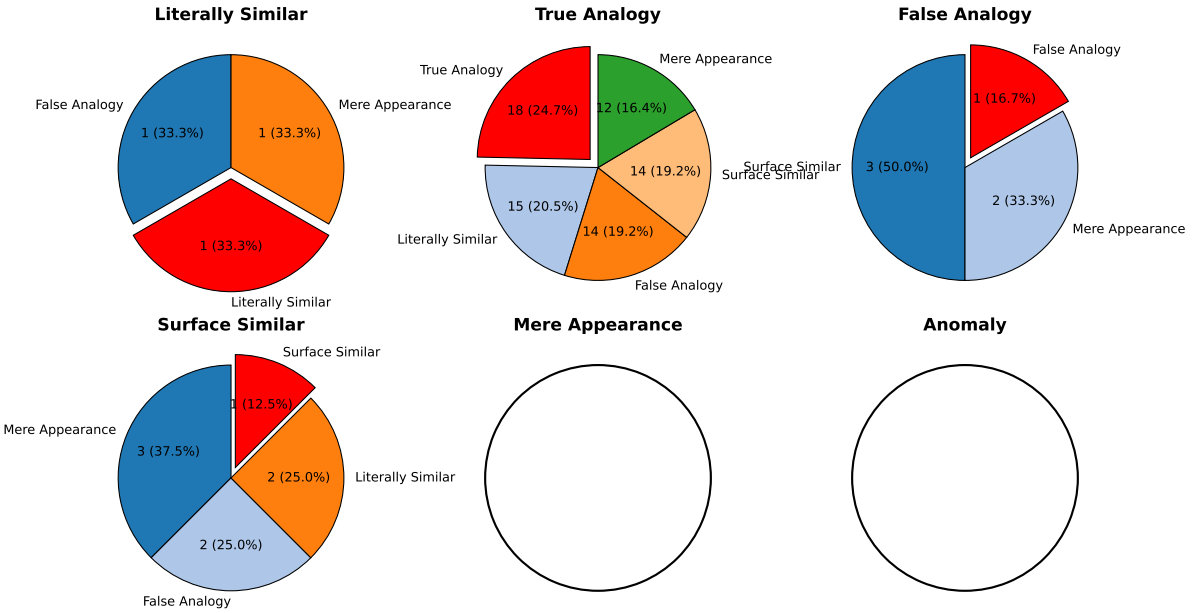

Figure 4: LLAMA3.1-8B(0 shot CoT)

LLAMA3.1-8B(1 shot CoT)

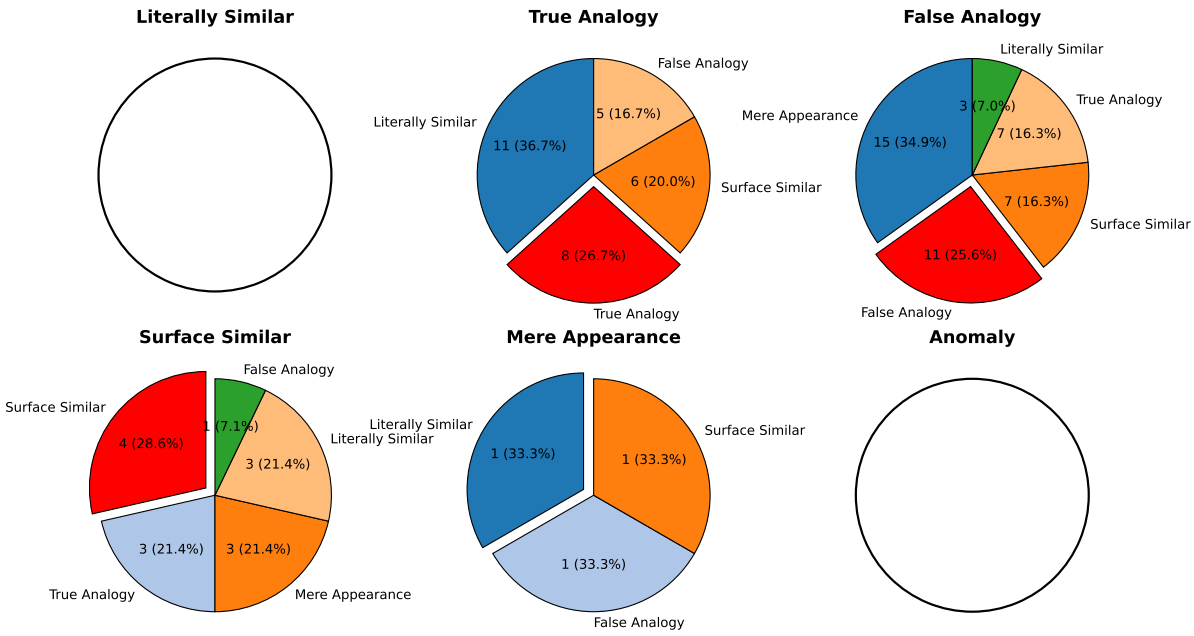

Figure 5: LLAMA3.1-8B(1 shot CoT)

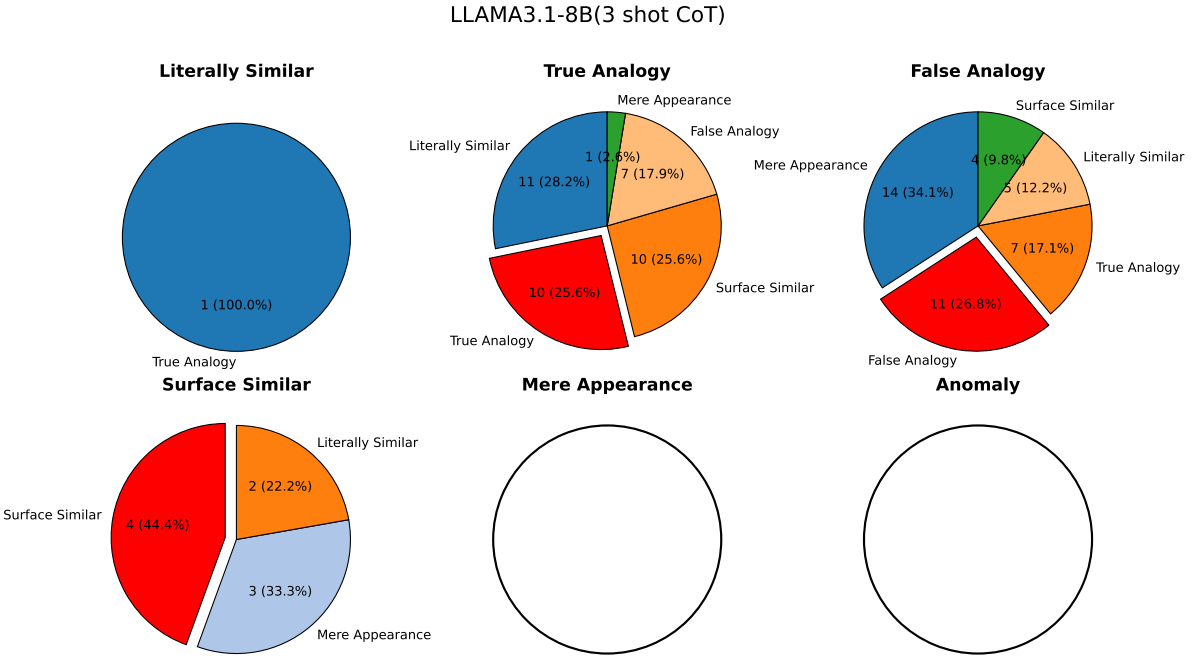

Figure 6: LLAMA3.1-8B(3 shot CoT)

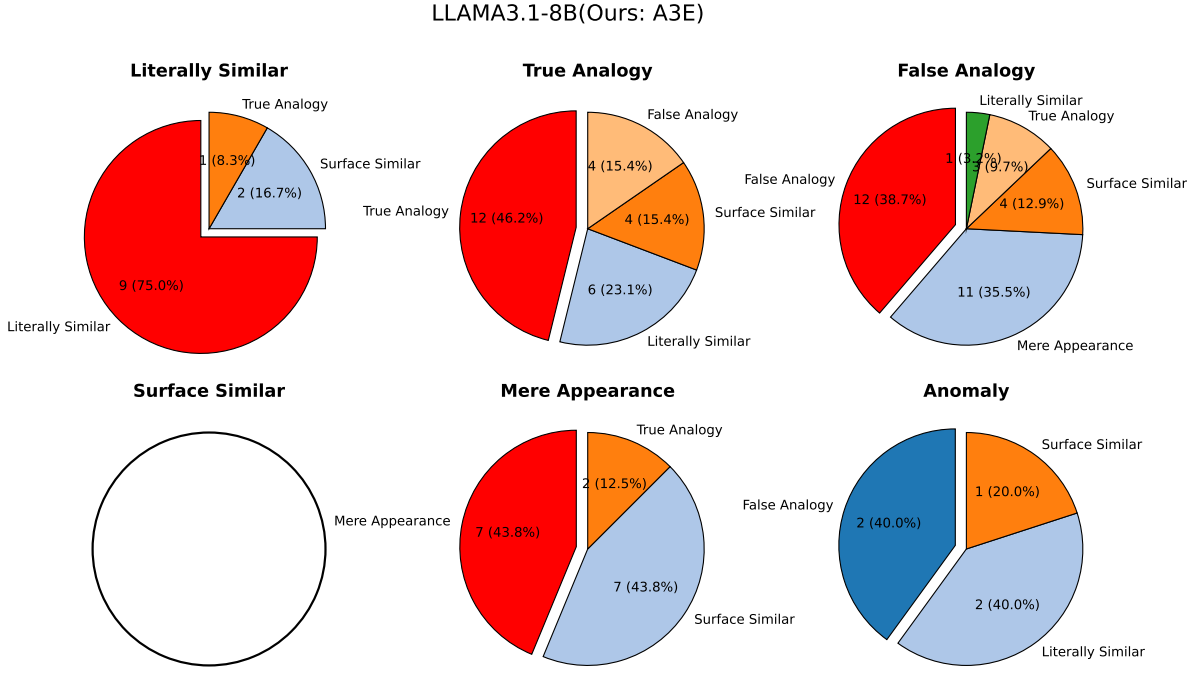

Figure 7: LLAMA3.1-8B-Ours-A3E

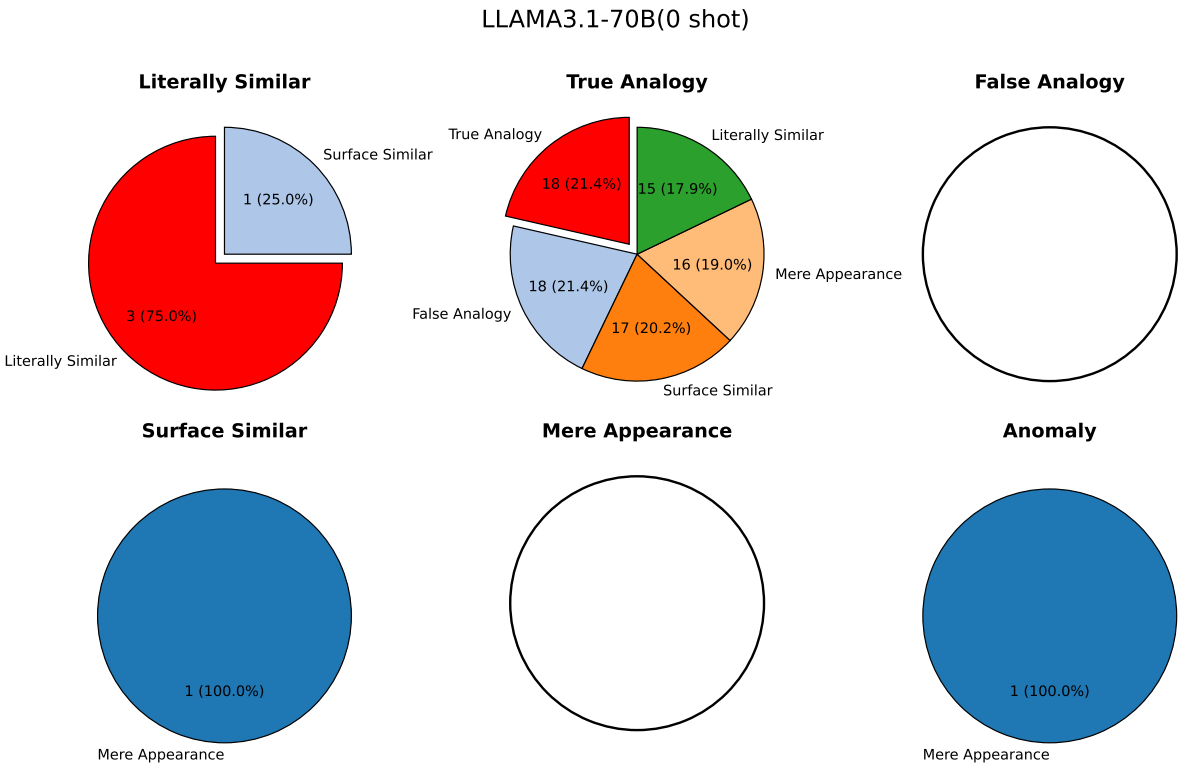

Figure 8: LLAMA3.1-70B(0 shot)

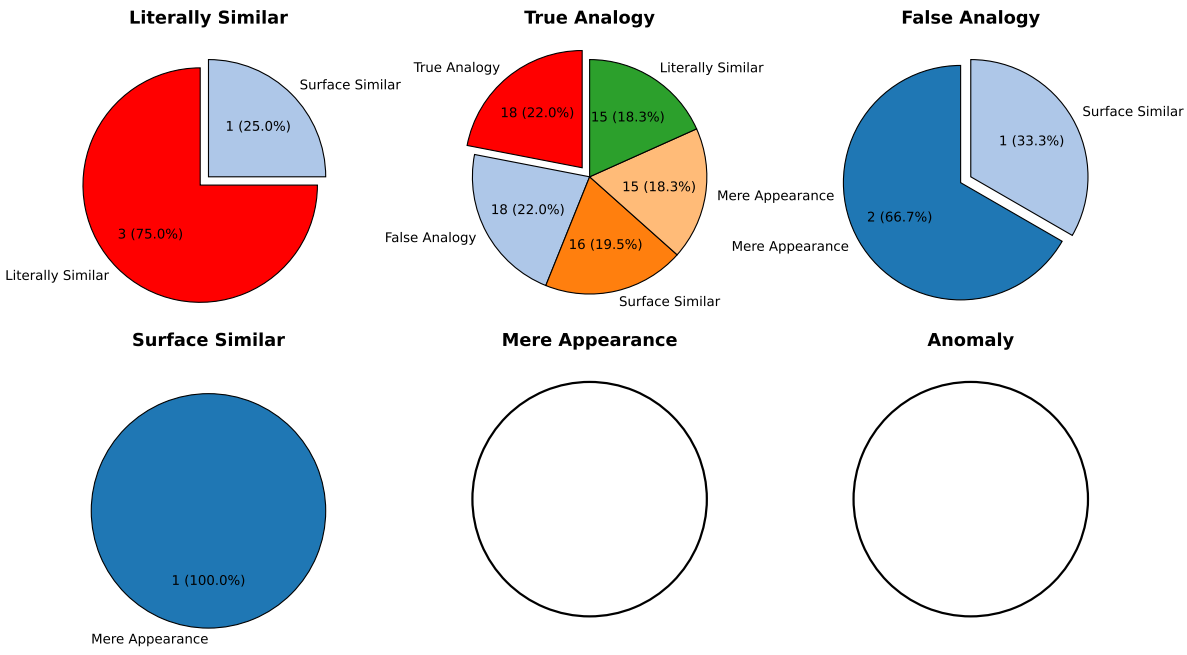

Figure 9: LLAMA3.1-70B(0 shot CoT)

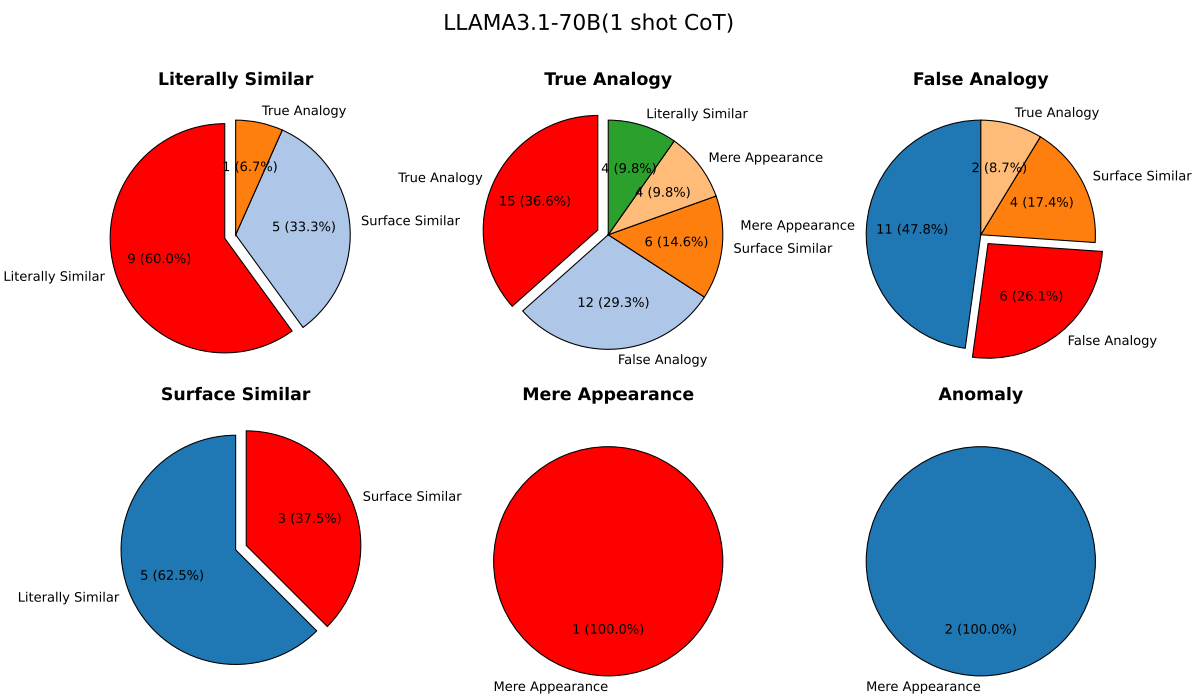

Figure 10: LLAMA3.1-70B(1 shot CoT)

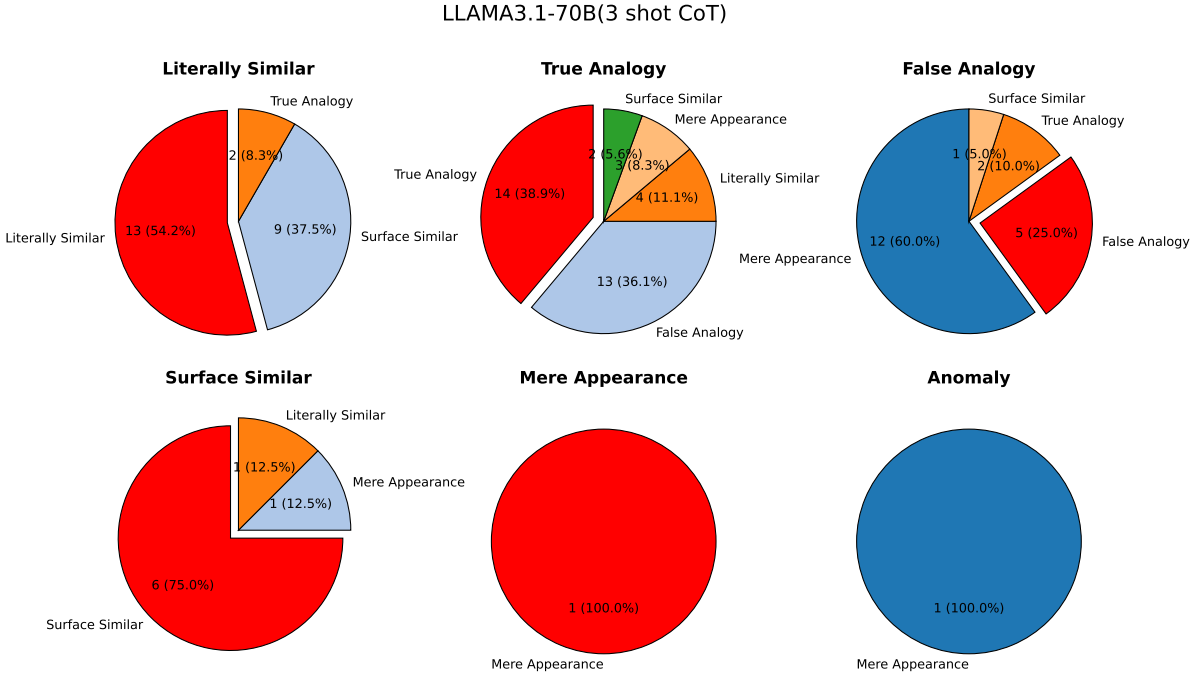

Figure 11: LLAMA3.1-70B(3 shot CoT)

LLAMA3.1-70B(Ours: A3E)

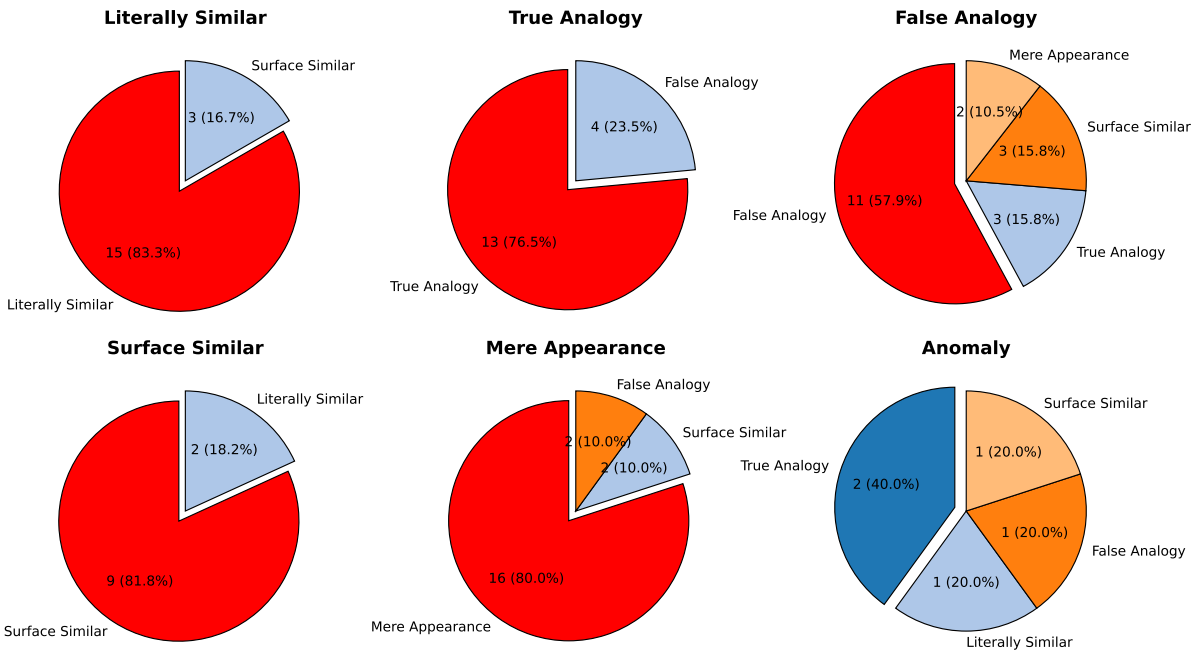

Figure 12: LLAMA3.1-70B-Ours-A3E

GPT-4o(0 shot)

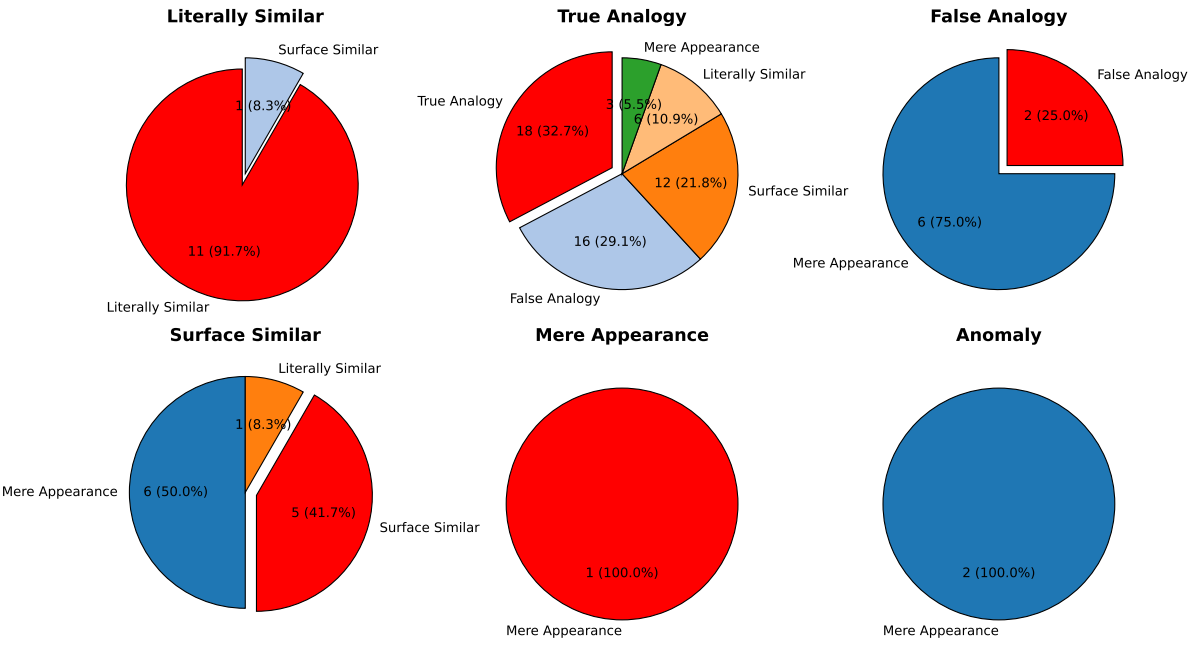

Figure 13: GPT-4o(0 shot)

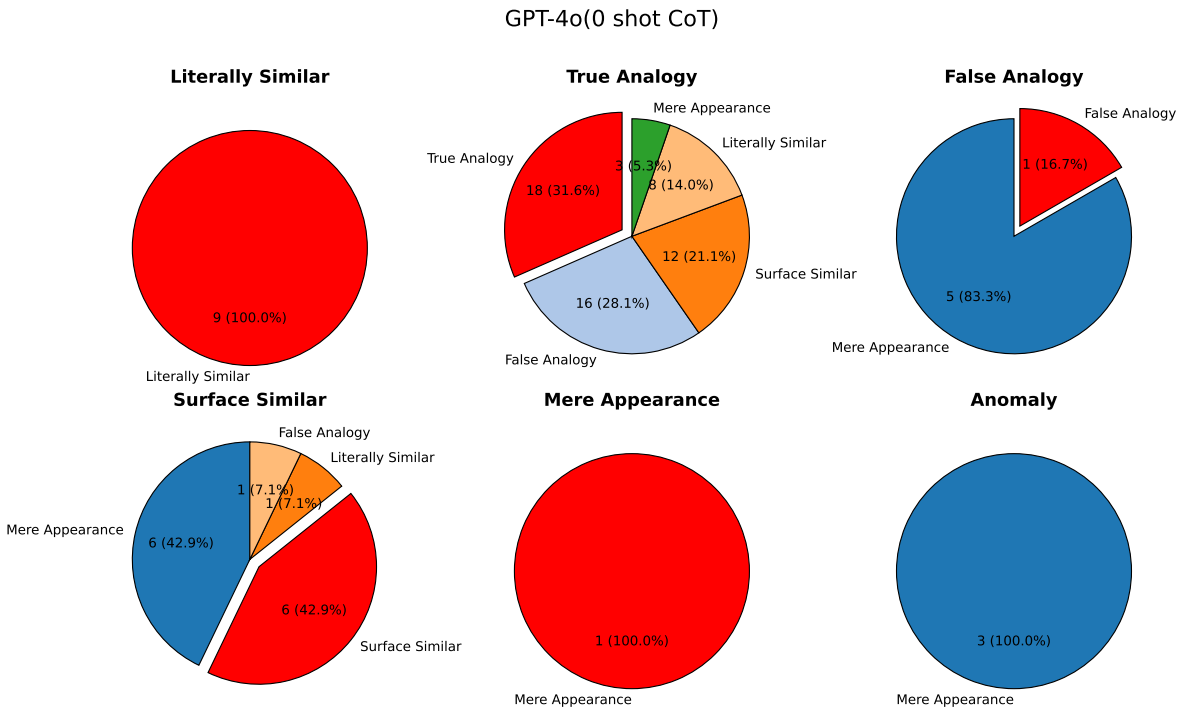

Figure 14: GPT-4o(0 shot CoT)

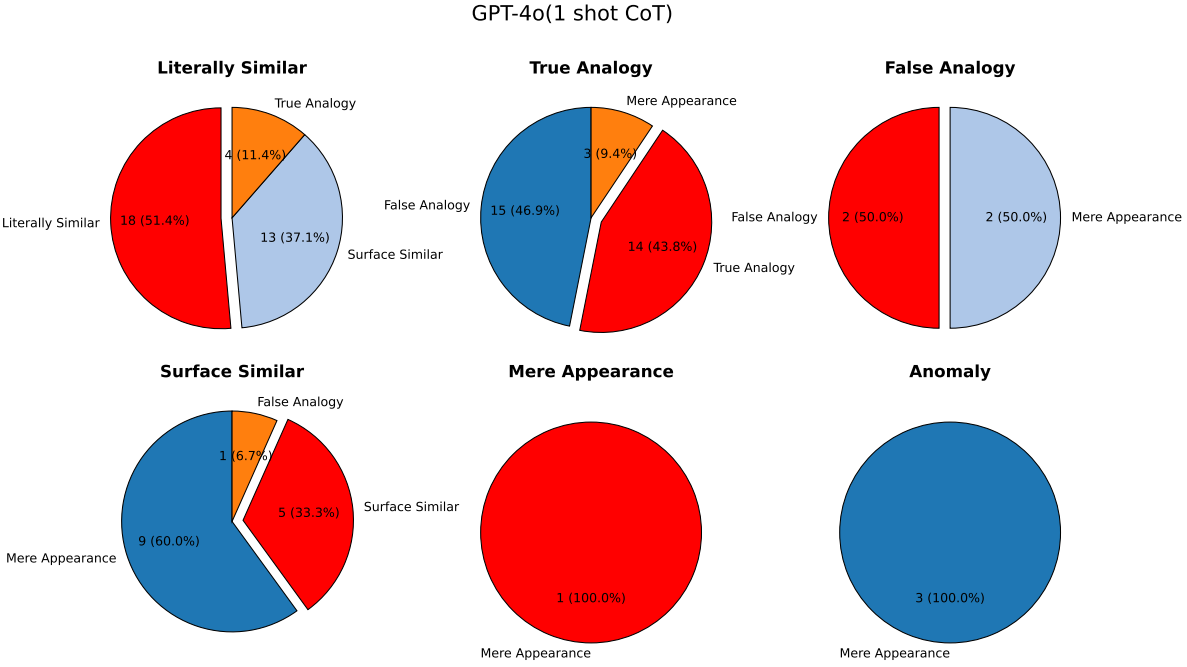

Figure 15: GPT-4o(1 shot CoT)

GPT-4o(3 shot CoT)

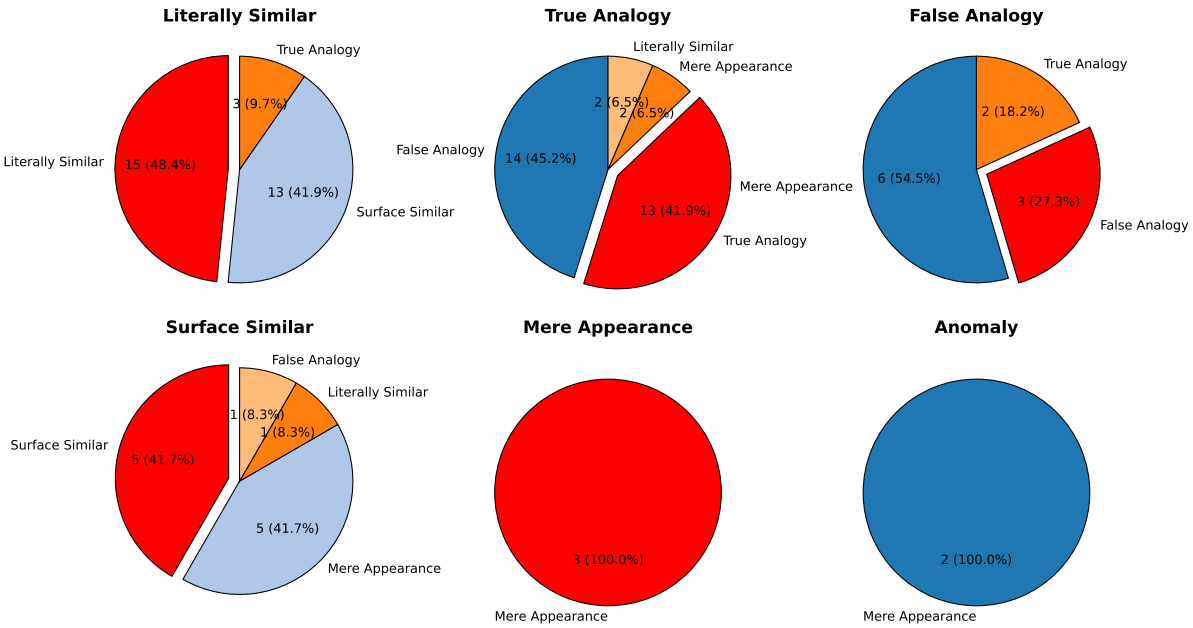

Figure 16: GPT-4o(3 shot CoT)

GPT-4o(Ours: A3E)

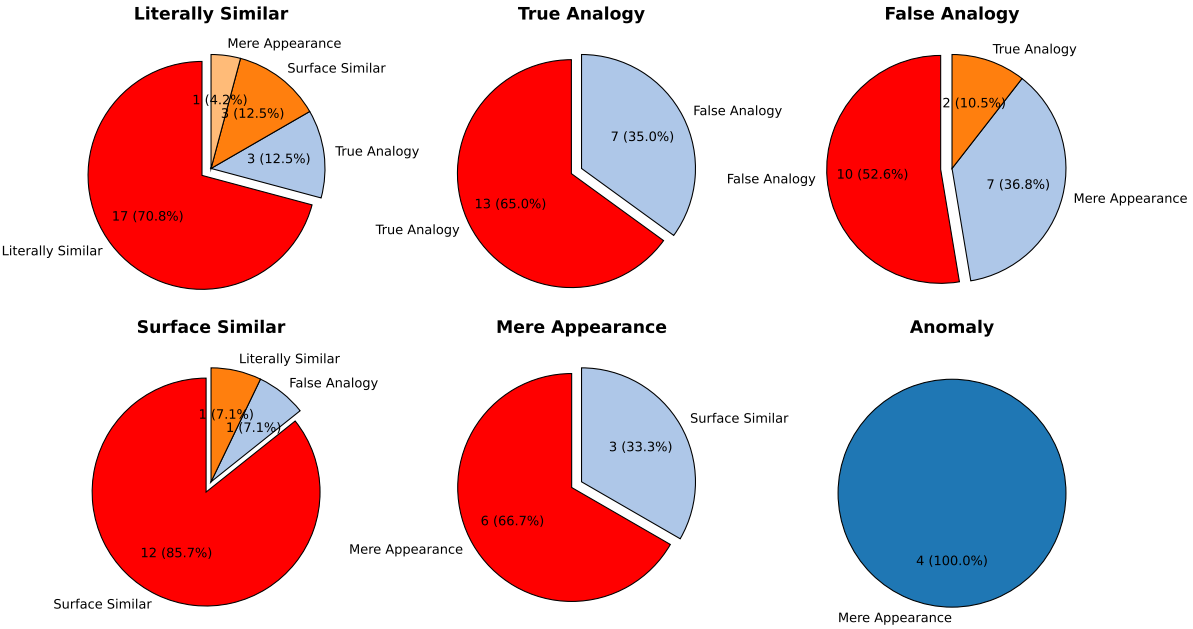

Figure 17: GPT-4o-Ours-A3E

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.

Case 1: Entity Analysis

Base: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.
Target: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.
Ground Truth: True Analogy

Answers:
From LLAMA3.1-8B:
- In the Base, the main characters are a hawk and a hunter. The hawk, aware of the hunter's intentions, offers to give the hunter some of its feathers, and the hunter agrees to never harm hawks again.
- In the Target, the main characters are a country called Zerdia and its warlike neighbor, Gagrach. Zerdia, aware of Gagrach's intentions, offers to sell some of its computers, and Gagrach agrees to never attack Zerdia again.
Therefore, specific background setting (a hawk and a hunter versus a country and its neighbor) are not similar: the context of the stories differ. The character roles and responsibilities (hawk and hunter, Zerdia and Gagrach) are also not similar: the hawk and hunter are in a transactional relationship, while Zerdia and Gagrach are in a conflict relationship. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a conflict leading to a peaceful resolution. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From LLAMA3.1-70B:
- In the Base, the main character is Karla, an old hawk, who lives in a tall oak tree and encounters a hunter who wants her feathers. The hawk offers the hunter some feathers, and the hunter agrees to never shoot at hawks again.
- In the Target, the main character is the country of Zerdia, which possesses the world's smartest computer and is attacked by its neighbor, Gagrach. Zerdia offers to sell some computers to Gagrach, and Gagrach promises to never attack Zerdia again.
Therefore, specific background setting (a hawk and a hunter versus a country and its neighbor) are different: a natural setting differs from a geopolitical setting. The character roles and responsibilities (hawk and hunter, country and neighbor) are also different: the hawk and hunter are in a predator-prey relationship, while the countries are in a diplomatic relationship. However, there is a certain similarity in the plot progression and dynamics, as both stories involve an initial conflict, a realization of the opponent's desire, and a resolution through offering what the opponent wants. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From GPT-4o:
- In the Base, the main characters are Karla, an old hawk, and a hunter. The hunter attempts to shoot the hawk but misses. The hawk offers her feathers to the hunter, who then pledges not to shoot hawks again.
- In the Target, the main characters are the country Zerdia and its warlike neighbor, Gagrach. Gagrach attacks Zerdia but fails. Zerdia offers to sell its computers to Gagrach, which then promises not to attack Zerdia again.
Therefore, the specific background settings (hawk and hunter versus countries in conflict) are different: an animal interaction differs from international relations. The character roles and responsibilities (hawk and hunter, Zerdia and Gagrach) are also different: the hawk and hunter are in a predator-prey relationship, while Zerdia and Gagrach are in a geopolitical conflict. However, there is a similarity in the plot progression and dynamics, as both stories involve an initial conflict that is resolved through a peaceful offering. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From Qwen2.5-72B:
- In the Base, the main characters are Karla, an old hawk, and a hunter. The hunter, aiming to get hawk feathers, misses his shot, and Karla offers him some feathers, leading to the hunter's promise to never shoot at a hawk again.
- In the Target, the main characters are the country of Zerdia and its neighbor, Gagrach. Gagrach, aiming to get Zerdian computers, launches a failed attack, and Zerdia offers to sell some computers, leading to Gagrach's promise to never attack Zerdia again.

Therefore, specific background settings (a hawk and a hunter versus two countries) are different: one is a natural setting, the other is a geopolitical setting. The character roles and responsibilities (hawk and hunter, countries and their governments) are also different: the hawk and hunter are in a predator-prey relationship, while the countries are in a conflict-resolution relationship. However, there is a certain similarity in the plot progression and dynamics, as both stories involve an initial conflict leading to a peaceful resolution. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

Case 1: Sentence Mapping

Base: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.
Target: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.
Ground Truth: True Analogy

Answers:
From LLAMA3.1-8B:
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.
//repeat prompt

1. Base: Karla, an old hawk, lived at the top of a tall oak tree. vs. Target: None.
2. Base: One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. vs. Target: None.
3. Base: The hunter took aim and shot at the hawk but missed. vs. Target: But the missiles were badly aimed and the attack failed.
4. Base: Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. vs. Target: The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country.
5. Base: The hunter was so grateful that he pledged never to shoot at a hawk again. vs. Target: The government of Gagrach was very pleased.
6. Base: He went off and shot deer instead. vs. Target: It promised never to attack Zerdia again.

From LLAMA3.1-70B:
1. Base: Karla, an old hawk, lived at the top of a tall oak tree. vs. Target: Once there was a small country called Zerdia that learned to make the world's smartest computer.
2. Base: One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. vs. Target: One day Zerdia was attacked by its warlike neighbor, Gagrach.
3. Base: The hunter took aim and shot at the hawk but missed. vs. Target: But the missiles were badly aimed and the attack failed.
4. Base: Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. vs. Target: The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country.

5. Base: The hunter was so grateful that he pledged never to shoot at a hawk again. vs. Target: The government of Gagrach was very pleased. It promised never to attack Zerdia again.
6. Base: He went off and shot deer instead. vs. Target: None.

From GPT-4o:
1. Base: Karla, an old hawk, lived at the top of a tall oak tree. vs. Target: Once there was a small country called Zerdia that learned to make the world's smartest computer.
2. Base: One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. vs. Target: One day Zerdia was attacked by its warlike neighbor, Gagrach.
3. Base: The hunter took aim and shot at the hawk but missed. vs. Target: But the missiles were badly aimed and the attack failed.
4. Base: Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. vs. Target: The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country.
5. Base: The hunter was so grateful that he pledged never to shoot at a hawk again. vs. Target: The government of Gagrach was very pleased.
6. Base: He went off and shot deer instead. vs. Target: It promised never to attack Zerdia again.

From Qwen2.5-72B:
1. Base: Karla, an old hawk, lived at the top of a tall oak tree. vs. Target: Once there was a small country called Zerdia that learned to make the world's smartest computer.
2. Base: One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. vs. Target: One day Zerdia was attacked by its warlike neighbor, Gagrach.
3. Base: The hunter took aim and shot at the hawk but missed. vs. Target: But the missiles were badly aimed and the attack failed.
4. Base: Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. vs. Target: The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country.
5. Base: The hunter was so grateful that he pledged never to shoot at a hawk again. vs. Target: The government of Gagrach was very pleased. It promised never to attack Zerdia again.
6. Base: He went off and shot deer instead. vs. Target: None.

Case 1: Relational Alignment

Base: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.
Target: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.
Ground Truth: True Analogy

Answers:
From LLAMA3.1-8B:
1. Base: Karla, an old hawk, lived at the top of a tall oak tree. vs. Target: None.
Classification: Irrelevant Group
2. Base: One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. vs. Target: None.
Classification: Irrelevant Group
3. Base: The hunter took aim and shot at the hawk but missed. vs. Target: But the missiles were badly aimed and the attack failed.
Classification: Similar Group
4. Base: Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. vs. Target: The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country.
Classification: Similar Group
5. Base: The hunter was so grateful that he pledged never to shoot at a hawk again. vs. Target: The government of Gagrach was very pleased.
Classification: Similar Group
6. Base: He went off and shot deer instead. vs. Target: It promised never to attack Zerdia again.
Classification: Similar Group
Summary:

```
Similar Group: [3, 4, 5, 6]
Dissimilar Group: []
Irrelevant Group: [1, 2]
```

From LLAMA3.1-70B:
1. In Base, it describes the hawk's living environment, while in Target, it describes the country's achievement. The roles described by the two are not corresponding in structure mapping, so it is classified as an irrelevant group.
2. In Base, the hawk observes the hunter, while in Target, the country is attacked by its neighbor. Both depict the main character being aware of a potential threat, so it is classified as a similar group.
3. In Base, the hunter's attack fails, while in Target, the missiles are badly aimed and the attack fails. Both describe the initial attack being unsuccessful, so it is classified as a similar group.
4. In Base, the hawk offers its feathers to the hunter, while in Target, the country offers to sell its computers to the neighbor. Both depict the main character offering something to the potential threat, so it is classified as a similar group.
5. In Base, the hunter is grateful and promises not to shoot at the hawk again, while in Target, the government is pleased and promises not to attack the country again. Both describe the potential threat being satisfied and making a promise, so it is classified as a similar group.
6. Because one side (Target) contains 'None', it is classified as an irrelevant group.
In summary,

```
Similar Group: [2, 3, 4, 5]
Dissimilar Group: []
Irrelevant Group: [1, 6]
```

From GPT-4o:
1. Both describe the initial setting and main character(s): Karla, an old hawk, and Zerdia, a small country. Both are classified as a similar group.
2. Both describe an antagonist preparing to attack: the hunter with a bow and arrows, and Gagrach attacking Zerdia. Both are classified as a similar group.
3. Both describe the antagonist's attack failing: the hunter missing the hawk, and Gagrach's missiles being badly aimed. Both are classified as a similar group.
4. Both describe the protagonist realizing the antagonist's intentions and offering a solution: Karla offering feathers, and Zerdia offering computers. Both are classified as a similar group.
5. Both describe the antagonist's reaction to the protagonist's offer: the hunter being grateful, and Gagrach being pleased. Both are classified as a similar group.
6. Both describe the antagonist's change in behavior: the hunter shooting deer instead, and Gagrach promising not to attack Zerdia again. Both are classified as a similar group.
In summary,

```
Similar Group: [1, 2, 3, 4, 5, 6]
Dissimilar Group: []
Irrelevant Group: []
```

From Qwen2.5-72B:
1. In the Base, Karla is an old hawk living in a tall oak tree, corresponding to Zerdia, a small country that learned to make the world's smartest computer. Both describe the main character or entity and their unique attribute, classified as a similar group.
2. In the Base, the hunter is seen with a bow and crude arrows, corresponding to Gagrach attacking Zerdia with missiles. Both describe an adversary with a weapon or attack, classified as a similar group.
3. In the Base, the hunter misses the hawk, corresponding to Gagrach's missiles failing to hit Zerdia. Both describe the failure of the attack, classified as a similar group.
4. In the Base, Karla offers feathers to the hunter, corresponding to Zerdia offering computers to Gagrach. Both describe the main character or entity offering something valuable to the adversary, classified as a similar group.
5. In the Base, the hunter pledges never to shoot at a hawk again, corresponding to Gagrach promising never to attack Zerdia again. Both describe the adversary making a promise to not repeat the hostile action, classified as a similar group.
6. Because one side (Target) contains 'None', it is classified as an irrelevant group.
In summary,

```
Similar Group: [1, 2, 3, 4, 5]
Dissimilar Group: []
Irrelevant Group: [6]
```

Case 1: Analogical Conclusion

Base: Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.
Target: Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.
Ground Truth: True Analogy

Answers:

From LLAMA3.1-8B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "4",
    "dissimilar-set count": "0",
    "irrelevant-set count": "2",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


From LLAMA3.1-70B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "4",
    "dissimilar-set count": "0",
    "irrelevant-set count": "2",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


From GPT-4o:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "6",
    "dissimilar-set count": "0",
    "irrelevant-set count": "0",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


From Qwen2.5-72B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "5",
    "dissimilar-set count": "0",
    "irrelevant-set count": "2",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


Case 2: Entity Analysis

Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. Finally, his cellmate got tired of this and demanded that he shut up. He should always play during the day, the cellmate insisted. The next night, Morris escaped from prison and hid in a boxcar. But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. Immediately, Morris was nabbed by the cops. Poor, simple-minded Morris wished he hadn't paid any

attention to his cellmate. Back in his cell, he devoted himself to covering the walls with obscenities.
Target: A little girl named Jenny wished she could run away from home. She loved to play with her dog every day, but only in her neighbor's yard. Eventually, her neighbor got fed-up and told her to always play in her own yard. But Jenny refused to listen because she thought it was bad luck to play in her own yard. The last time she did it her mother caught her right away and made her go to work.
Ground Truth: False Analogy

Answers:
From LLAMA3.1-8B:
- In the Base, the main characters are a prisoner named Morris and his cellmate. Morris, who loves to play the harmonica, escapes from prison but is caught again after playing his harmonica in a boxcar. He then returns to his cell and covers the walls with obscenities.
- In the Target, the main characters are a little girl named Jenny and her neighbor. Jenny wants to play with her dog in her neighbor's yard, but her neighbor tells her to play in her own yard. Jenny refuses to listen and gets caught by her mother.
Therefore, specific background setting (a prisoner in a cell versus a little girl at home) are not similar: the settings differ significantly. The character roles and responsibilities (prisoner and cellmate, little girl and neighbor) are also not similar: the prisoner and cellmate are in a competitive relationship, while the little girl and neighbor are in a relationship where the neighbor is giving instructions. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a character being told to do something and then doing it, leading to negative consequences. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From LLAMA3.1-70B:
- In the Base, the main character is a prisoner named Morris who loves to play the harmonica. His cellmate demands that he play during the day, and Morris's obedience to this rule ultimately leads to his recapture after escaping.
- In the Target, the main character is a little girl named Jenny who loves to play with her dog. Her neighbor tells her to play in her own yard, but Jenny disobeys, believing it's bad luck, and gets caught by her mother.
The specific background settings (prison life versus a child's play) are different: one is a confined environment, while the other is a domestic setting. The character roles and responsibilities (prisoner and cellmate, child and neighbor) are also different: the prisoner and cellmate are in a confined relationship, while the child and neighbor are in a advisory relationship. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a character's obedience or disobedience to advice leading to an undesirable outcome. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From GPT-4o:
- In the Base, the main characters are Morris, a simple-minded prisoner who loves to play the harmonica, and his cellmate who gets tired of the noise. Morris escapes but gets caught because he plays the harmonica as his cellmate suggested.
- In the Target, the main characters are Jenny, a little girl who loves to play with her dog, and her neighbor who gets fed up with her playing in his yard. Jenny refuses to play in her own yard due to superstition and gets caught by her mother.
Therefore, specific background settings (prison versus home/neighborhood) are different: prison life differs from a child's home life. The character roles and responsibilities (prisoner and cellmate, girl and neighbor) are also different: the prisoner and cellmate have a conflict over noise, while the girl and neighbor have a conflict over play location. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a character ignoring advice and facing consequences. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

From Qwen2.5-72B:
- In the Base, the main characters are Morris, a simple-minded prisoner, and his cellmate. Morris is often playing the harmonica at night, which annoys his cellmate. After escaping, Morris follows his cellmate's advice to play during the day, leading to his capture. Back in prison, he vents his frustration by covering the walls with obscenities.

- In the Target, the main characters are Jenny, a little girl, and her neighbor. Jenny loves to play with her dog in her neighbor's yard, which annoys the neighbor. Despite the neighbor's advice to play in her own yard, Jenny refuses, believing it is bad luck. Eventually, her mother catches her and makes her go to work.
Therefore, specific background settings (prison escape versus a child playing in a neighbor's yard) are different: one is a criminal scenario, the other a domestic situation. The character roles and responsibilities (prisoner and cellmate, child and neighbor) are also different: the prisoner and cellmate are in a confined, adversarial relationship, while the child and neighbor are in a community setting. However, there is a certain similarity in the plot progression and dynamics, as both stories involve a character ignoring or following advice, leading to a negative outcome. Additionally, there are no common words.
In summary,

```
{
   "background": "False",
   "role": "False",
   "plot": "True",
   "common-words count": []
}
```

Case 2: Sentence Mapping

Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. Finally, his cellmate got tired of this and demanded that he shut up. He should always play during the day, the cellmate insisted. The next night, Morris escaped from prison and hid in a boxcar. But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. Immediately, Morris was nabbed by the cops. Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. Back in his cell, he devoted himself to covering the walls with obscenities.
Target: A little girl named Jenny wished she could run away from home. She loved to play with her dog every day, but only in her neighbor's yard. Eventually, her neighbor got fed-up and told her to always play in her own yard. But Jenny refused to listen because she thought it was bad luck to play in her own yard. The last time she did it her mother caught her right away and made her go to work.
Ground Truth: False Analogy

Answers:
From LLAMA3.1-8B:
1. Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. vs. Target: A little girl named Jenny wished she could run away from home.
2. Base: Finally, his cellmate got tired of this and demanded that he shut up. vs. Target: She loved to play with her dog every day, but only in her neighbor's yard.
3. Base: He should always play during the day, the cellmate insisted. vs. Target: Eventually, her neighbor got fed-up and told her to always play in her own yard.
4. Base: The next night, Morris escaped from prison and hid in a boxcar. vs. Target: But Jenny refused to listen because she thought it was bad luck to play in her own yard.
5. Base: But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. vs. Target: The last time she did it her mother caught her right away and made her go to work.
6. Base: Immediately, Morris was nabbed by the cops. vs. Target: None.
7. Base: Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. vs. Target: None.
8. Base: Back in his cell, he devoted himself to covering the walls with obscenities. vs. Target: None.

From LLAMA3.1-70B:
1. Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. vs. Target: A little girl named Jenny wished she could run away from home.
2. Base: Finally, his cellmate got tired of this and demanded that he shut up. vs. Target: Eventually, her neighbor got fed-up and told her to always play in her own yard.
3. Base: He should always play during the day, the cellmate insisted. vs. Target: None.
4. Base: The next night, Morris escaped from prison and hid in a boxcar. vs. Target: None.
5. Base: But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. vs. Target: But Jenny refused to listen because she thought it was bad luck to play in her own yard.
6. Base: Immediately, Morris was nabbed by the cops. vs. Target: The last time she did it her mother caught her right away and made her go to work.
7. Base: Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. vs. Target: None.
8. Base: Back in his cell, he devoted himself to covering the walls with obscenities. vs. Target: None.

From GPT-4o:
1. Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. vs. Target: A little girl named Jenny wished she could run away from home.
2. Base: Finally, his cellmate got tired of this and demanded that he shut up. vs. Target: She loved to play with her dog every day, but only in her neighbor's yard.
3. Base: He should always play during the day, the cellmate insisted. vs. Target: Eventually, her neighbor got fed-up and told her to always play in her own yard.
4. Base: The next night, Morris escaped from prison and hid in a boxcar. vs. Target: But Jenny refused to listen because

she thought it was bad luck to play in her own yard.
5. Base: But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. vs. Target: The last time she did it her mother caught her right away and made her go to work.
6. Base: Immediately, Morris was nabbed by the cops. vs. Target: None.
7. Base: Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. vs. Target: None.
8. Base: Back in his cell, he devoted himself to covering the walls with obscenities. vs. Target: None.

From Qwen2.5-72B:
1. Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. vs. Target: A little girl named Jenny wished she could run away from home.
2. Base: Finally, his cellmate got tired of this and demanded that he shut up. vs. Target: She loved to play with her dog every day, but only in her neighbor's yard.
3. Base: He should always play during the day, the cellmate insisted. vs. Target: Eventually, her neighbor got fed-up and told her to always play in her own yard.
4. Base: The next night, Morris escaped from prison and hid in a boxcar. vs. Target: But Jenny refused to listen because she thought it was bad luck to play in her own yard.
5. Base: But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. vs. Target: The last time she did it her mother caught her right away and made her go to work.
6. Base: Immediately, Morris was nabbed by the cops. vs. Target: None.
7. Base: Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. vs. Target: None.
8. Base: Back in his cell, he devoted himself to covering the walls with obscenities. vs. Target: None.

Case 2: Relational Alignment

Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. Finally, his cellmate got tired of this and demanded that he shut up. He should always play during the day, the cellmate insisted. The next night, Morris escaped from prison and hid in a boxcar. But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. Immediately, Morris was nabbed by the cops. Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. Back in his cell, he devoted himself to covering the walls with obscenities.
Target: A little girl named Jenny wished she could run away from home. She loved to play with her dog every day, but only in her neighbor's yard. Eventually, her neighbor got fed-up and told her to always play in her own yard. But Jenny refused to listen because she thought it was bad luck to play in her own yard. The last time she did it her mother caught her right away and made her go to work.
Ground Truth: False Analogy

Answers:
From LLAMA3.1-8B:
1. Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. vs. Target: A little girl named Jenny wished she could run away from home.
Analysis: Both describe the main character having a desire or tendency that is not aligned with the expectations of others. Morris wants to play the harmonica at night, while Jenny wants to play in her neighbor's yard. This is classified as a similar group.
2. Base: Finally, his cellmate got tired of this and demanded that he shut up. vs. Target: She loved to play with her dog every day, but only in her neighbor's yard.
Analysis: The cellmate's demand is a reaction to Morris's harmonica playing, while the neighbor's demand is a reaction to Jenny playing in their yard. However, the cellmate's demand is a direct cause of Morris's change in behavior, whereas the neighbor's demand is not directly related to Jenny's behavior. This is classified as an irrelevant group.
3. Base: He should always play during the day, the cellmate insisted. vs. Target: Eventually, her neighbor got fed-up and told her to always play in her own yard.
Analysis: Both describe a character giving advice or making a demand that is not directly related to the main character's behavior. However, the cellmate's advice is a direct cause of Morris's change in behavior, while the neighbor's demand is not directly related to Jenny's behavior. This is classified as an irrelevant group.
4. Base: The next night, Morris escaped from prison and hid in a boxcar. vs. Target: But Jenny refused to listen because she thought it was bad luck to play in her own yard.
Analysis: Morris's escape is a direct result of his desire to play the harmonica, while Jenny's refusal to listen is a result of her own superstition. However, both characters are making decisions based on their own desires or beliefs, which is classified as a similar group.
5. Base: But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. vs. Target: The last time she did it her mother caught her right away and made her go to work.
Analysis: Morris's decision to play the harmonica is a direct result of his cellmate's advice, while Jenny's punishment is a result of her own disobedience. However, both characters are facing consequences for their actions, which is classified as a similar group.
6. Base: Immediately, Morris was nabbed by the cops. vs. Target: None.
Analysis: This is classified as an irrelevant group because there is no corresponding statement in the Target.
7. Base: Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. vs. Target: None.
Analysis: This is classified as an irrelevant group because there is no corresponding statement in the Target.
8. Base: Back in his cell, he devoted himself to covering the walls with obscenities. vs. Target: None.
Analysis: This is classified as an irrelevant group because there is no corresponding statement in the Target.

Summary:

```
Similar Group: [1, 4, 5]
Dissimilar Group: []
Irrelevant Group: [2, 3, 6, 7, 8]
```

From LLAMA3.1-70B:
1. Because one side (Target) does not describe the main character's hobby or behavior, it is classified as an irrelevant group.
2. In Base, the cellmate is fed up with Morris's nighttime harmonica playing, while in Target, the neighbor is fed up with Jenny playing in their yard. Both describe a character getting tired of the main character's behavior, classified as a similar group.
3. Because one side (Target) contains 'None', it is classified as an irrelevant group.
4. Because one side (Target) contains 'None', it is classified as an irrelevant group.
5. In Base, Morris follows his cellmate's advice and plays the harmonica during the day, while in Target, Jenny refuses to follow her neighbor's advice to play in her own yard. Both describe the main character's decision regarding someone else's advice, but the actions are opposite, so it is classified as a dissimilar group.
6. In Base, Morris is caught by the police after playing the harmonica, while in Target, Jenny is caught by her mother after playing in her own yard. Both describe the main character getting caught after making a decision, classified as a similar group.
7. Because one side (Target) contains 'None', it is classified as an irrelevant group.
8. Because one side (Target) contains 'None', it is classified as an irrelevant group.
In summary,

```
Similar Group: [2, 6]
Dissimilar Group: [5]
Irrelevant Group: [1, 3, 4, 7, 8]
```

From GPT-4o:
1. Because one side (Target) contains 'None', it is classified as an irrelevant group.
2. In the Base, Morris's cellmate gets tired of his harmonica playing and demands he stop, corresponding to the Target where Jenny's neighbor gets fed up with her playing in their yard. Both describe someone getting annoyed and demanding a change in behavior, classified as a similar group.
3. In the Base, the cellmate insists Morris should play during the day, while in the Target, the neighbor insists Jenny should play in her own yard. Both involve someone imposing a restriction on the protagonist's behavior, classified as a similar group.
4. In the Base, Morris escapes and hides, while in the Target, Jenny refuses to listen because of her belief. Both describe the protagonists taking actions based on their beliefs or decisions, classified as a similar group.
5. In the Base, Morris remembers his cellmate's advice and starts playing the harmonica, leading to his capture. In the Target, Jenny's previous experience of being caught by her mother when playing in her own yard is mentioned. Both describe the protagonists facing consequences due to their actions, classified as a similar group.
6. Because one side (Target) contains 'None', it is classified as an irrelevant group.
7. Because one side (Target) contains 'None', it is classified as an irrelevant group.
8. Because one side (Target) contains 'None', it is classified as an irrelevant group.
In summary,

```
Similar Group: [2, 3, 4, 5]
Dissimilar Group: []
Irrelevant Group: [1, 6, 7, 8]
```

From Qwen2.5-72B:
1. Because one side (Target) describes a wish to run away, while the other (Base) describes a habitual action, they are not directly aligned in structure. Therefore, it is classified as an irrelevant group.
2. In the Base, the cellmate gets tired of Morris playing the harmonica at night, while in the Target, the neighbor gets tired of Jenny playing with her dog in the neighbor's yard. Both describe a character getting annoyed by another's repetitive action, classified as a similar group.
3. In the Base, the cellmate insists Morris should play during the day, while in the Target, the neighbor tells Jenny to play in her own yard. Both describe a character giving a directive to change the behavior, classified as a similar group.
4. In the Base, Morris escapes from prison and hides in a boxcar, while in the Target, Jenny refuses to listen to the neighbor because she believes it's bad luck. The actions and reasons are different, so it is classified as a dissimilar group.
5. In the Base, Morris starts playing the harmonica because he remembers the cellmate's directive, while in the Target, Jenny is caught by her mother for playing in the neighbor's yard. Both describe the main character facing consequences due to following or not following a directive, but the specific actions and outcomes are different, so it is classified as a dissimilar group.
6. Because one side (Target) contains 'None', it is classified as an irrelevant group.
7. Because one side (Target) contains 'None', it is classified as an irrelevant group.
8. Because one side (Target) contains 'None', it is classified as an irrelevant group.
In summary,

```
Similar Group: [2, 3]
Dissimilar Group: [4, 5]
Irrelevant Group: [1, 6, 7, 8]
```

Case 2: Analogical Conclusion

Base: Morris was a simple-minded prisoner who loved to play the harmonica at night. Finally, his cellmate got tired of this and demanded that he shut up. He should always play during the day, the cellmate insisted. The next night, Morris escaped from prison and hid in a boxcar. But then Morris remembered that his cellmate said he should play harmonica every day, so he started playing. Immediately, Morris was nabbed by the cops. Poor, simple-minded Morris wished he hadn't paid any attention to his cellmate. Back in his cell, he devoted himself to covering the walls with obscenities.
Target: A little girl named Jenny wished she could run away from home. She loved to play with her dog every day, but only in her neighbor's yard. Eventually, her neighbor got fed-up and told her to always play in her own yard. But Jenny refused to listen because she thought it was bad luck to play in her own yard. The last time she did it her mother caught her right away and made her go to work.
Ground Truth: False Analogy

Answers:
From LLAMA3.1-8B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "3",
    "dissimilar-set count": "0",
    "irrelevant-set count": "5",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


From LLAMA3.1-70B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "2",
    "dissimilar-set count": "1",
    "irrelevant-set count": "5",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "dissimilar",
    "predict_label": "False Analogy"
}
```


From GPT-4o:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "4",
    "dissimilar-set count": "0",
    "irrelevant-set count": "4",
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "similar",
    "predict_label": "True Analogy"
}
```


From Qwen2.5-72B:


```
{
    "background": "False",
    "role": "False",
    "plot": "True",
    "common-words count": "0",
    "similar-set count": "2",
    "dissimilar-set count": "2",
    "irrelevant-set count": "4",
```

```
    "entities": "dissimilar",
    "one-order relations": "similar",
    "higher-order relations": "dissimilar",
    "predict_label": "False Analogy"
}
```