Analogical Retrieval

Jixiao Zhang
jzhan432@jh.edu

Chunsheng Zuo
czuo3@jh.edu

Abstract

Analogical retrieval, the task of identifying passages that share abstract relational similarities with a query, presents unique challenges that traditional dense retrieval models often fail to address. Unlike semantic similarity, analogical reasoning requires recognizing deep relational structures, making it critical for applications in storytelling, multi-step reasoning, and creative problem-solving. In this work, we propose and evaluate methods to enhance analogical retrieval performance through a combination of external augmentation and prompt optimization.

First, we employ GPT-4 in a few-shot setup to generate new S1-like summaries from S10 data, providing insights into the effectiveness of AI-generated summaries (*Summary_FS*) for analogical reasoning tasks. Second, we systematically analyze prompt effectiveness by categorizing prompts into *Helpful*, *Tricky/Harmful*, and *Irrelevant*, identifying configurations that maximize performance for relational reasoning. Third, we evaluate various summarization strategies, including *Summary*, *+Summary*, and *Summary_FS*, across lightweight and advanced retrieval models. Our results highlight the benefits of summary augmentation, the utility of prompt optimization, and the contrasting preferences of retrieval models, with Promptriever excelling across most configurations.

1 Introduction

In the realm of natural language processing, analogical retrieval presents a unique challenge. Unlike traditional dense retrieval tasks, which focus on matching lexically or semantically similar content, analogical retrieval aims to connect passages that share a deeper, abstract similarity. For instance, given a query describing a situation involving poetic justice, an analogical retrieval system should prioritize passages with analogous themes of consequence and irony over those sharing only superficial keywords. This abstract form of reasoning—identifying connections based on shared structures rather than surface details—remains a complex task for standard retrieval models.

Analogical reasoning is essential for AI agents tasked with solving problems that require a nuanced understanding of relationships between concepts, stories, or scenarios. This form of reasoning supports applications across diverse fields, from enhancing creative processes in storytelling and science to aiding multi-step question answering where relationships between intermediate steps mirror broader analogies. However, existing dense retrieval methods, typically trained on semantic similarity, often overlook these nuanced relationships, resulting in mismatches when retrieving analogically relevant information.

This project aims to bridge this gap by exploring methods that adapt retrieval models to recognize analogical relationships effectively. Specifically, we propose and evaluate the following contributions:

- **Few-Shot Summarization for Analogical Retrieval:** We employ GPT-4 to generate new S1-like summaries from existing S10 data using a few-shot learning setup. These generated summaries (*Summary_FS*) are evaluated for their effectiveness in improving analogical retrieval on S10, providing insights into the utility of few-shot learning for summary generation.
- **Prompt Optimization for Relational Reasoning:** We systematically analyze the impact of diverse prompt types on the performance of Promptriever, categorizing prompts into *Helpful*, *Tricky/Harmful*, and *Irrelevant* to identify the optimal configurations for analogical retrieval.
- **Comparison of Summarization Strategies:** We assess the effectiveness of dif-

ferent summarization techniques, including *Summary* (human-crafted), *+Summary* (augmented queries), and *Summary_FS* (few-shot GPT-4 summaries), across various retrieval models, highlighting the alignment between model architecture and summarization strategies.

- **Benchmarking Analogical Retrieval Models:** We provide a comprehensive evaluation of lightweight and advanced retrieval models, including Paraphrase-MiniLM-L12-v2, All-MPNET-Base-v2, Sentence-T5-XL, and Promptriever, under diverse task configurations, datasets, and summarization techniques.

Through these contributions, our work aims to enhance the understanding of analogical retrieval and inform future research into aligning retrieval models with abstract and relational reasoning tasks.

2 Literature Review

2.1 Analogical Matching Methods

For analogical retrieval, two main approaches emerge: manually designed algorithms and large language model-based analysis. These methods address the need for relational abstraction, which is critical for effective analogy retrieval.

2.1.1 Manually Designed Algorithms

Some analogical retrieval methods rely on structured, manual designs that parse passages for entities, roles, and relationships, applying these features to construct a relational representation. Sultan and Shahaf (2022) proposed a system where sentences are parsed into entities with roles using QA-SRL (Question Answer-driven Semantic Role Labeling) and then matched across passages by mapping entities using the cosine similarity of sentence-BERT embeddings. Similarly, Kang et al. (2022) explored entity-role identification in an analogical context by training RNN models to identify purpose-related words, further improving analogy detection.

2.1.2 Large Language Model (LLM)-Based Analysis

Large language models (LLMs) have demonstrated significant potential in performing abstract reasoning and analogical matching. (Webb et al., 2023) showed that LLMs like GPT-4 exhibit emergent analogical reasoning capabilities, particularly for tasks requiring verbal or story-based analogy detection. AnaloBench, introduced by (Ye et al., 2024), presents a benchmark specifically for testing the analogical reasoning capabilities of LLMs across various lengths of passages, showing that larger LLMs can perform analogical reasoning, although performance declines for longer passages. This benchmark highlights the need for enhanced abstractions in LLMs to address the challenges of analogical retrieval for extended contexts.

2.2 Augmentation Strategies for Analogical Retrieval

There are ongoing efforts to enhance analogical retrieval performance by augmenting query and passage inputs with key elements, entity relationships, or summary information, aiming to improve the retrieval model's ability to capture analogical structures. (Roth et al., 2024) developed a retrieval-augmented generation (RAG) system that rewrites queries by extracting essential conceptual nouns that represent each sentence. By focusing on key elements in retrieval, Roth's approach offers an effective strategy for capturing analogical relationships, making it a valuable augmentation technique for analogical retrieval tasks. Similarly, models like ANALOGYKB (Yuan et al., 2024) focus on augmenting relational understanding for analogical reasoning by building a large-scale knowledge base of analogy pairs, structured to enhance the analogical reasoning capabilities of LLMs. ANALOGYKB includes resources that identify relations both of the same type and of analogous but distinct types, addressing a gap in traditional dense retrieval models by expanding the relational diversity in embeddings.

2.3 Fine-Tuning Retrieval Models for Analogy-Focused Embeddings

Beyond external augmentation, analogical retrieval can also benefit from fine-tuning retrieval models to produce embeddings optimized specifically for analogy recognition. Internal methods such as Low-Rank Adaptation modules (LoRA) (Hu et al., 2021; Dettmers et al., 2023) offers a modular approach for fine-tuning. This could enable models to incorporate multiple analogy-focused features with each being LoRA fine-tuned independently.

We would also take reference from recent progress in retriever fine-tuning. For example, RankLLaMA(Ma et al., 2023) leverages LLaMA-based models in a multi-stage retrieval pipeline, com-

bining hard negatives from dense and sparse retrievals to improve ranking accuracy. Similarly, the Augmentation-Adapted Retriever (Yu et al., 2023) serves as a plug-in for various LMs, using cross-attention to identify relevant information without task-specific retraining. We would also look at (Lin et al., 2023), which has explored a variety of augmentation strategies to improve the robustness of the fine-tuned retriever.

3 Methods

3.1 Task Description

The task of analogical retrieval involves selecting the most analogous option to a given query from four candidates. We evaluate our methods on two benchmark tasks from the Analobench dataset:

- **T1S1**: Sentence-level analogies, requiring matching of structural and contextual relationships at the sentence granularity.
- **T1S10**: Story-level analogies, which involve reasoning over more complex and abstract relationships.

3.2 Query Expansion

We enrich queries using summarization techniques to provide models with varying levels of contextual information. Summaries are generated at three levels:

- **Keyword**: A single keyword representing the query's main concept.
- **Keyphrase**: A concise phrase capturing the query's key idea.
- **Summary**: A full sentence providing a detailed summary of the query.

The summaries can either augment the query (*+Keyword*, *+Keyphrase*, *+Summary*) or replace the query entirely (*Summary Only*). This setup allows us to analyze the impact of varying levels of summarization on retrieval performance.

The summaries are generated using the GPT-4 API from OpenAI with a carefully designed prompt aimed at extracting all possible abstractions, morals, and key ideas from a given sentence or story. The prompt instructs GPT-4 to summarize the input in three distinct forms—*Keyword*, *Keyphrase*, and *Full Summary Sentence*—while focusing on universal themes, abstract concepts, and generalized relationships. By explicitly avoiding specific details (e.g., names or places), the summaries emphasize relational and thematic elements relevant for analogical reasoning.

Each generated summary is structured into three components: a single-word *Keyword*, a concise *Keyphrase*, and a comprehensive *Summary Sentence*, separated by semicolons. For example:

> *"Deception; Deception and Consequences; An individual's persuasive deception led a group to misplace their trust, resulting in disillusionment and loss when the truth was revealed."*

To evaluate their effectiveness, each type of summary (*Keyword*, *Keyphrase*, and *Summary Sentence*) was tested independently, either augmenting the query (*+Keyword*, *+Keyphrase*, *+Summary*) or replacing it entirely (*Summary Only*). This approach enables a detailed analysis of how different levels of abstraction contribute to retrieval performance, without conflating the effects of multiple summary types.

3.3 Few-Shot Learning with GPT-4-Generated Data

To explore the potential of few-shot learning for generating S1-like summaries, we utilized GPT-4 to create new data from the Analogy Benchmark (S10). Using the existing data as prompts, GPT-4 was instructed to generate concise summaries capturing the relational and contextual essence of each story. These summaries are referred to as *Summary_FS* (few-shot summaries).

The models were evaluated under the following configurations:

- **Summary**: Original summaries done in Section 3.2.
- **+Summary**: Queries augmented with the original summaries.
- **Summary_FS**: Few-shot summaries generated by GPT-4.
- **+Summary_FS**: Queries augmented with the few-shot summaries.

Performance was measured on S10, with the primary metric being accuracy (percentage of correctly identified analogous stories). This setup allows us to compare the effectiveness of the diverse zero-shot-generated summary and to the few-shot summaries in improving analogical reasoning.

3.4 Models

We evaluate five retrieval models, each representing a different approach to encoding and retrieving text:

- **Paraphrase-MiniLM-L12-v2**: A lightweight embedding model optimized for paraphrase identification and semantic similarity.
- **All-MiniLM-L12-v1**: Another lightweight embedding model, trained for general-purpose text similarity tasks.
- **All-MPNET-Base-v2**: A slightly larger model designed to improve sentence embeddings and capture nuanced relationships.
- **Sentence-T5-XL**: A powerful transformer-based model designed for robust encoding of sentences and paragraphs.
- **Promptriever**: A model leveraging prompt-based techniques for analogical retrieval, fine-tuned to focus on relational and abstract reasoning over semantic similarity.

3.5 Evaluation Framework

Candidate options for each query are ranked based on similarity scores, with specific evaluation techniques tailored to the models used:

- **Cosine Similarity**: Used to compare embeddings of queries and options. This method measures similarity by calculating the cosine of the angle between vectors in the embedding space.
- **Prompt-Based Relational Reasoning**: For Promptriever, carefully designed prompts guide the model to focus on relational structures, emphasizing analogical reasoning over surface-level semantic similarity.

Prompts were evaluated across three categories: *Helpful*, *Tricky/Harmful*, and *Irrelevant*, to examine their impact on analogical retrieval. Each category reflects varying levels of clarity and alignment with the task objectives. Accuracy was used as the primary evaluation metric, measuring the proportion of queries where the top-ranked candidate matches the ground truth.

3.6 Experimental Setup

We conducted experiments on both S1 and S10 tasks, evaluating the impact of summarization techniques and prompt categories on retrieval performance. The evaluation included the following configurations:

- **Query Only**: The query is presented without any summarization.
- **+Keyword, +Keyphrase, +Summary**: The query is augmented with one of the respective summaries, providing varying levels of contextual enrichment.
- **Summary Only**: The query is replaced entirely by a summary.

Prompts were tested in all configurations to analyze their interaction with summarization strategies. For each combination, the performance of five retrieval models (Paraphrase-MiniLM-L12-v2, All-MiniLM-L12-v1, All-MPNET-Base-v2, Sentence-T5-XL, and Promptriever) was assessed. This comprehensive design allows for a systematic comparison of summarization techniques, prompt effectiveness, and model performance across tasks of differing complexity.

3.7 Implementation Details

The experiments were conducted on a GPU-enabled environment to support efficient inference for large transformer-based models. Each model was evaluated with appropriate batch sizes to balance computational efficiency and memory usage. Preprocessing steps ensured consistency in formatting queries and options across tasks.

4 Results: Model and Query Expansion Performance

4.1 Effectiveness of Summarization Techniques

Summarization techniques significantly impact model performance across both S1 and S10 tasks, with *Summary* emerging as the most effective approach overall. This section analyzes the performance improvements brought by summarization techniques in detail.

General Observations. Full sentence summaries (*Summary*) consistently provided substantial performance improvements across all models, particularly for lightweight architectures. On S1, *Sum-*

Model	Query	Keyword	Keyphrase	Summary	+Keyword	+Keyphrase	+Summary
Paraphrase-MiniLM-L12-v2	57.4	61.2	63.2	66.8	62.9	62.6	63.8
All-MiniLM-L12-v1	51.2	60.9	67.1	65.9	62.6	61.8	62.4
All-MPNET-Base-v2	56.2	62.4	62.9	67.4	59.1	58.5	66.2
Sentence-T5-XL	72.6	65.6	71.2	77.9	76.2	80.6	78.2
Promptriever	70.3	56.2	68.5	75.3	71.2	75.9	79.1

Table 1: Performance on S1 (Accuracy across summarization techniques). Here, *Query* represents no summarization; *+Keyword*, *+Keyphrase*, and *+Summary* represent the combination of query with the respective summarization technique.

Model	Query	Keyword	Keyphrase	Summary	+Keyword	+Keyphrase	+Summary
Paraphrase-MiniLM-L12-v2	38.8	43.5	43.8	53.8	43.5	42.9	42.4
All-MiniLM-L12-v1	40.0	44.1	45.9	51.2	45.6	45.3	44.1
All-MPNET-Base-v2	42.4	42.6	47.4	53.8	49.7	50.6	51.2
Sentence-T5-XL	54.4	50.3	50.3	56.5	55.6	53.5	53.2
Promptriever	59.7	41.2	46.8	55.0	61.2	60.0	60.3

Table 2: Performance on S10 (Accuracy across summarization techniques). Here, *Query* represents no summarization; *+Keyword*, *+Keyphrase*, and *+Summary* represent the combination of query with the respective summarization technique.

Model	Dataset	None Accuracy	Best Technique	Best Technique Accuracy	Improvement
Paraphrase-MiniLM-L12-v2	S1	57.4	Summary	66.8	9.4
All-MiniLM-L12-v1	S1	51.2	Keyphrase	67.1	15.9
All-MPNET-Base-v2	S1	56.2	Summary	67.4	11.2
Sentence-T5-XL	S1	72.6	Query + Keyphrase	80.6	8.0
Promptriever	S1	70.3	Query + Summary	79.1	8.8

Table 3: Best summarization technique for each model on S1.

Model	Dataset	None Accuracy	Best Technique	Best Technique Accuracy	Improvement
Paraphrase-MiniLM-L12-v2	S10	38.8	Summary	53.8	15.0
All-MiniLM-L12-v1	S10	40.0	Summary	51.2	11.2
All-MPNET-Base-v2	S10	42.4	Summary	53.8	11.4
Sentence-T5-XL	S10	54.4	Summary	56.5	2.1
Promptriever	S10	59.7	Query + Keyword	61.2	1.5

Table 4: Best summarization technique for each model on S10.

mary boosted the accuracy of Paraphrase-MiniLM-L12-v2 from 57.4 to 66.8 and All-MPNET-Base-v2 from 56.2 to 67.4. Similarly, on S10, *Summary* improved Paraphrase-MiniLM-L12-v2 from 38.8 to 53.8 and All-MPNET-Base-v2 from 42.4 to 53.8, marking it as the best-performing technique for these models.

Advanced models such as Sentence-T5-XL also benefited from *Summary*, achieving strong results on both datasets. On S1, Sentence-T5-XL achieved 77.9 with *Summary*, an improvement of 5.3 over its performance with *Query*. On S10, Sentence-T5-XL improved from 54.4 to 56.5 with *Summary*, although the improvement was modest compared to lightweight models.

Dataset-Specific Trends. Performance trends differed significantly between S1 and S10. On S1, models exhibited larger performance gains with *Summary*, likely due to the dataset's relatively straightforward analogical reasoning. For example, All-MiniLM-L12-v1 achieved an accuracy of 65.9 with *Summary*, an improvement of 14.7 over its performance with *Query*. On S10, however, the gains from *Summary* were more modest for some models, particularly advanced ones. Promptriever, for instance, achieved its best performance (61.2) on S10 with *+Keyword*, improving only slightly over its performance with *Query*.

Key Insight. These results highlight that *Summary* is a reliable technique across diverse model architectures and datasets, particularly excelling on tasks that do not involve overly complex reasoning. Lightweight models derive the greatest benefit from *Summary*, as the richer contextual information compensates for their limited reasoning capacity.

4.2 Model-Specific Preferences for Summarization

Models exhibited distinct preferences for summarization techniques, with advanced models favoring combined strategies, while lightweight models performed best with standalone techniques such as *Keyphrase* or *Summary*.

Advanced Models. Advanced models such as Sentence-T5-XL and Promptriever adapted well to richer and combined techniques. On S1, Sentence-T5-XL achieved the highest accuracy (80.6) with +*Keyphrase*, an improvement of 8.0 over its performance with *Query*. Promptriever, similarly, performed best with +*Summary*, achieving 79.1, an improvement of 8.8 over its performance with *Query*. These results suggest that advanced models can effectively integrate contextual and relational signals when provided with richer inputs.

However, on S10, the benefits of combined techniques diminished. Promptriever achieved its best performance (61.2) with +*Keyword*, improving only slightly over its performance with *Query*. Sentence-T5-XL, despite achieving a modest improvement of 2.1 with *Summary*, did not exhibit the same level of adaptability as seen on S1. This indicates that the increased complexity of S10 limits the effectiveness of richer or combined strategies.

Lightweight Models. Lightweight models, including Paraphrase-MiniLM-L12-v2, All-MiniLM-L12-v1, and All-MPNET-Base-v2, performed best with simpler standalone techniques. On S1, All-MiniLM-L12-v1 achieved the highest accuracy improvement among all models (15.9) with *Keyphrase*, increasing from 51.2 to 67.1. Paraphrase-MiniLM-L12-v2 and All-MPNET-Base-v2 also demonstrated strong gains with *Summary*, improving by 9.4 and 11.2, respectively. These results reflect the alignment of lightweight models with simpler summarization strategies that do not overwhelm their capacity for reasoning.

On S10, lightweight models continued to favor *Summary*. Paraphrase-MiniLM-L12-v2 improved by 15.0, the largest gain observed on S10, while All-MPNET-Base-v2 improved by 11.4. In contrast, combined strategies did not provide any additional benefits for these models, as seen with All-MiniLM-L12-v1, which performed best with *Summary* (11.2 improvement).

Key Insight. The contrasting preferences of advanced and lightweight models underline the importance of tailoring summarization strategies to the model architecture. Advanced models benefit from the richer context and relational signals provided by combined techniques but struggle with overly complex datasets. Lightweight models, on the other hand, excel with concise and standalone summarization strategies that align with their limited capacity.

5 Results: Prompt Influence

5.1 Effectiveness of Prompts by Category

Prompts categorized as *Helpful* generally outperformed those in the *Tricky/Harmful* and *Irrelevant* categories, especially in the +*Summary* configuration. For example, the prompt "A relevant document would be the most analogous to the query..." achieved 79.1 on S1 and 60.3 on S10 with +*Summary*, highlighting the importance of clear and targeted guidance. By contrast, *Tricky/Harmful* prompts such as "A relevant document demonstrates a relational analogy..." scored only 72.1 on S1 and 52.4 on S10 under the same configuration, showing limited benefits even with summary augmentation.

5.2 Impact of "Irrelevant" Prompts

Interestingly, prompts categorized as *Irrelevant*, such as "Merry Christmas" and "No Prompt," showed relatively strong performance in the *Query* configuration but plateaued or underperformed in +*Summary*. For instance, "No Prompt" achieved 65.6 on S1 and 56.8 on S10 in the *Query* configuration but offered minimal gains with summaries, reaching only 77.6 on S1 and 59.7 on S10 with +*Summary*. This suggests that while models can perform reasonably well without explicit guidance, they benefit more from well-crafted prompts.

Prompt (Full Text)	Category
"A relevant document demonstrates a relational analogy to the query, focusing on parallels in context, structure, or reasoning rather than direct semantic overlap. Ensure that the documents adhere to these criteria by avoiding those that diverge into tangential or overly literal interpretations. Additionally, exclude passages from [specific field/domain] unless they offer clear analogical insights."	Tricky/Harmful
"A relevant document would be the least analogous to the query."	Tricky/Harmful
"A relevant document would be the least semantically similar to the query. I care about analogical similarity."	Tricky/Harmful
"Merry Christmas."	Irrelevant
"No Prompt"	Irrelevant
"A relevant document would be the most analogous to the query. A relevant document would be the most analogous to the query. I don't care about semantic similarity. I don't care about semantic similarity."	Helpful
"A relevant document would be the most analogous to the query. I don't care about semantic similarity."	Helpful
"A relevant document would be the most analogous to the query. I don't care about semantic similarity. A relevant document would be the most analogous to the query. I don't care about semantic similarity."	Helpful
"Focus on high-level concepts, abstraction, and key ideas."	Helpful

Table 5: List of prompts used in the experiments and their corresponding categories.

Prompt	Category	S1		S10	
		Query	+Summary	Query	+Summary
Prompt 1	Tricky/Harmful	49.1	72.1	52.6	52.4
Prompt 2	Tricky/Harmful	57.9	78.2	52.9	57.9
Prompt 3	Tricky/Harmful	58.8	77.1	52.4	57.9
Prompt 4	Irrelevant	64.7	78.2	57.4	58.2
Prompt 5	Irrelevant	65.6	77.6	56.8	59.7
Prompt 6	Helpful	70.3	**79.1**	59.7	**60.3**
Prompt 7	Helpful	67.9	78.2	59.1	58.2
Prompt 8	Helpful	**70.3**	**79.1**	**60.0**	59.7
Prompt 9	Helpful	54.4	77.4	55.6	58.5

Table 6: Experimental results for different prompts tested with Promptriever. S1 and S10 columns group results for the *Query* and *+Summary* configurations. Prompts are referred to by abbreviated names (*Prompt 1*, *Prompt 2*, etc.), corresponding to their full descriptions in Table 5. The highest value in each column is highlighted in bold.

5.3 Performance of "Tricky/Harmful" Prompts

Prompts deemed *Tricky/Harmful* generally underperformed across both configurations. For example, "A relevant document demonstrates a relational analogy..." achieved only 49.1 on S1 and 52.6 on S10 in the *Query* configuration. Even with *+Summary*, this prompt failed to meaningfully improve performance, achieving 72.1 on S1 and 52.4 on S10. These results highlight the negative impact of ambiguous or overly verbose prompts, which can mislead the model.

5.4 Task-Level Performance Trends

Across all prompts, S1 consistently achieved higher scores than S10, reflecting the relative simplicity of sentence-level analogies compared to the more abstract reasoning required for story-level analogies. For instance, the prompt "A relevant document would be the least semantically similar to the query..." scored 77.1 on S1 but only 57.9 on S10 in the *+Summary* configuration. This performance gap emphasizes the challenges posed by the increased complexity of S10.

5.5 Effectiveness of *+Summary*

The *+Summary* configuration provided significant improvements for most *Helpful* prompts. For example, the prompt "A relevant document would be the most analogous to the query..." improved from 70.3 (*Query*, S1) to 79.1 (*+Summary*, S1). However, for *Tricky/Harmful* and *Irrelevant* prompts, the gains were limited or negligible. For instance, "Merry Christmas" improved from 64.7 (*Query*, S1) to only 78.2 (*+Summary*, S1), with similar stagnation observed for S10.

5.6 Clarity and Repetition in Prompts

Prompts with clear and repeated guidance performed best, particularly in the *+Summary* configuration. For example, repeated phrases in "A relevant document would be the most analogous to the query..." consistently achieved strong results, with 79.1 on S1 and 59.7 on S10. Similarly, simple prompts like "Focus on high-level concepts, abstraction, and key ideas." performed well on S1 (77.4), indicating that clarity is crucial for guiding the model effectively.

5.7 Key Takeaways

- **Helpful Prompts Excel**: Clear and concise prompts significantly boost performance, particularly with summary augmentation.
- **Irrelevant Prompts Are Surprisingly Effective**: While *Irrelevant* prompts perform well in the *Query* configuration, they fail to leverage the benefits of query expansion with *+Summary*.
- **Tricky Prompts Hinder Performance**: Ambiguous or overly verbose prompts negatively impact the model, even with additional context from summaries.
- **Task Complexity Matters**: The performance gap between S1 and S10 highlights the challenges of abstract reasoning in story-level analogies.
- **Clarity and Repetition Improve Results**: Repeated or explicit prompts provide the best guidance, ensuring the model focuses on relevant analogical relationships.

6 Results: Feshow Summary

7 Results: Few-Shot Summary Effects

7.1 Impact of Few-Shot Summaries on Model Performance

The use of GPT-4-generated summaries (*Summary_FS*) provided valuable insights into the effectiveness of few-shot learning for analogical reasoning. This section analyzes model performance across different configurations.

General Observations. The models demonstrated varied performance when using *Summary_FS* compared to the GPT zero-shot generated summaries (*Summary*). Promptriever achieved the highest accuracy across all configurations, demonstrating its adaptability to both zero-shot and few-shot generated summaries. For example, Promptriever scored 55.0% with *Summary* and 51.2% with *Summary_FS*. However, its performance improved significantly with augmentation, achieving 60.3% with *+Summary* and 59.4% with *+Summary_FS*.

Comparison of Summarization Techniques. The zero-shot generated summaries (*Summary*) generally outperformed the GPT-4-generated few-shot summaries (*Summary_FS*) in standalone configurations:

- Sentence-T5-XL scored 56.5% with *Summary* compared to 51.5% with *Summary_FS*.
- Paraphrase-MiniLM-L12-v2 achieved 53.8% with *Summary*, while its performance dropped to 40.6% with *Summary_FS*.

This suggests that the zero-shot generated summaries better captured abstract analogical relationships compared to the few-shot summaries.

Effectiveness of Augmented Configurations. When queries were augmented with summaries (*+Summary* and *+Summary_FS*), performance improved across most models. For example:

Model	Summary (%)	+Summary (%)	Summary_FS (%)	+Summary_FS (%)
All-MiniLM-L12-v1	51.2	44.1	46.2	44.1
All-MPNET-Base-v2	53.8	51.2	42.1	43.8
Paraphrase-MiniLM-L12-v2	53.8	42.4	40.6	42.1
Promptriever	55.0	60.3	51.2	59.4
Sentence-T5-XL	56.5	53.2	51.5	55.0

Table 7: Experimental results using few-shot learning with GPT-4-generated data. Models were evaluated on S10 using two types of summaries: *Summary* (original S1-like summaries) and *Summary_FS* (generated few-shot summaries). +*Summary* and +*Summary_FS* denote the combination of the query with the respective summaries. All values are reported as percentages.

- Promptriever achieved 60.3% with +*Summary* and 59.4% with +*Summary_FS*.
- Sentence-T5-XL also showed an improvement, scoring 53.2% with +*Summary* and 55.0% with +*Summary_FS*.

These results highlight the utility of summary augmentation, regardless of whether the summaries were generated using zero-shot or few-shot approaches.

7.2 Model-Specific Trends

The models exhibited distinct preferences for the type of summaries used:

- **Promptriever**: Consistently performed best across all configurations, with minimal performance degradation when using *Summary_FS* instead of *Summary*.
- **Lightweight Models**: Models such as All-MiniLM-L12-v1 and Paraphrase-MiniLM-L12-v2 showed significant drops in performance when using *Summary_FS*, indicating a reliance on higher-quality zero-shot summaries.
- **All-MPNET-Base-v2**: Demonstrated balanced performance across all configurations, scoring 53.8% with *Summary* and 51.2% with +*Summary_FS*.

7.3 Key Insights

- **Zero-Shot Summaries Excel**: GPT zero-shot generated summaries (*Summary*) consistently outperformed few-shot GPT-4-generated summaries (*Summary_FS*) in standalone configurations.
- **Augmentation Boosts Performance**: Both +*Summary* and +*Summary_FS* configurations improved accuracy across models, demonstrating the utility of enriched query inputs.
- **Model-Specific Preferences**: Advanced models like Promptriever adapted well to few-shot summaries, while lightweight models struggled, highlighting the importance of summary quality.
- **Few-Shot Potential**: Despite underperforming compared to zero-shot summaries, GPT-4-generated few-shot summaries showed promise, particularly when used in augmented configurations.

8 Conclusion

This work addresses the unique challenges of analogical retrieval by proposing methods to improve the alignment of retrieval models with abstract relational reasoning tasks. We explored the potential of few-shot learning with GPT-4 to generate new summaries (*Summary_FS*) and evaluated their effectiveness in analogical retrieval on S10. Our findings indicate that while human-crafted summaries (*Summary*) remain superior, GPT-4-generated summaries show promise, particularly when used in augmented configurations (+*Summary_FS*).

We also systematically analyzed the impact of diverse prompt strategies on retrieval performance, categorizing prompts into *Helpful*, *Tricky/Harmful*, and *Irrelevant*. This analysis highlighted the importance of prompt clarity and repetition in enhancing relational reasoning, with *Helpful* prompts achieving the highest performance.

Furthermore, our evaluation of summarization techniques across lightweight and advanced retrieval models revealed distinct preferences. Lightweight models benefitted most from standalone strategies like *Summary*, while advanced

models excelled with richer, augmented configurations such as +*Summary*. Among the models, Promptriever demonstrated robust performance across all setups, showcasing its adaptability to diverse configurations and summarization strategies.

These findings contribute to advancing the understanding of analogical retrieval and underscore the need for tailored approaches that balance model capabilities with task complexity. Future work may explore integrating additional relational reasoning frameworks or expanding the scope to more diverse datasets, further enhancing the capabilities of analogical retrieval systems.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 29(6):1–36.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. *arXiv preprint arXiv:2302.07452*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

K Roth, Rushil Gupta, Simon Halle, and Bang Liu. 2024. Pairing analogy-augmented generation with procedural memory for procedural q&a. *arXiv preprint arXiv:2409.01344*.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. *arXiv preprint arXiv:2210.12197*.

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

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*.

Dongxu Yu, Jeremy Mallen, et al. 2023. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

A Zero-Shot Prompt for Summary Generation

The following prompt in Figure 1 was used to generate GPT zero-shot summaries.

B Few-Shot Prompt for Summary Generation

The following prompt in Figure 2 was used to generate GPT few-shot summaries(Convert the story back to one sentence).

(*Summary*) for the analogical l retrieval task.

Zero-Shot Prompt for GPT-4 Summary Generation

Please read the following story carefully. Your task is to identify and summarize all of its key ideas, capturing the significant underlying principles, themes, or morals. Focus on abstract concepts and relationships. In your summaries:

List All Significant Key Ideas: - Extract multiple key ideas if applicable. - Ensure each key idea represents a distinct theme, principle, or moral from the story.

Use Generalized Language: - Avoid specific details such as names, places, or unique objects. - Refer to characters and entities in general terms (e.g., "an individual," "a group," "a community").

Highlight Core Concepts and Relationships: - Focus on the main events, actions, and outcomes in broad terms. - Describe cause-and-effect relationships and the dynamics between characters or forces.

Emphasize Underlying Themes or Morals: - Include universal themes such as justice, irony, betrayal, redemption, self-discovery, consequences of actions, etc. - Consider both explicit messages and implicit lessons.

Be Clear and Concise: - Keep each key idea brief but comprehensive enough to convey its essence. - Ensure that each summary stands alone and makes sense without additional context.

Example Output: If the story involves someone deceiving others and then being deceived themselves, but also touches on themes of karma and loss, your summaries might be: - "Deception; Deception and Consequences; An individual's persuasive deception led a group to misplace their trust, resulting in disillusionment and loss when the truth was revealed." - "Illusion; The Power of Illusion; Superficial appearances can easily mislead those who rely solely on what they see, highlighting the importance of critical examination." - "Greed; The Role of Greed; Desire for wealth and material gain can cloud judgment, making individuals vulnerable to exploitation."

Output Format: - First comes the key word of the summary, then a key phrase, and finally the full summary sentence, separated by semicolons (";"). - Output each summary on its own line. - Avoid adding any extra symbols like "*" or quotation marks.

Story Input: Here is the story: {story}

Figure 1

Few-Shot Prompt for GPT-4 Summary Generation

Task: Convert the detailed story into a concise, single sentence that encapsulates its essential idea, story content or moral.
Here are some examples: Sentence: "A smile can hide a thousand tears."
Story: "Sophia was the life of every party, her laughter lighting up rooms wherever she went. Her friends adored her wit, and she was always the one to comfort others when they were down. What no one knew was that she cried herself to sleep most nights, mourning the loss of her father. At work, she excelled, masking her pain with perfectionism, but the strain began to show when she stopped returning friends' calls. One day, her coworker Mia found her sitting quietly in the break room, tears streaming down her face. Mia gently encouraged her to talk, and for the first time in years, Sophia opened up. She began attending grief counseling and journaling her feelings. Slowly, her laughter returned—not the forced kind, but genuine joy. Her friends noticed the change and rallied around her. Sophia learned that healing doesn't mean forgetting, and she could honor her father's memory by living fully."
Sentence: "The ancient clock tower in the square has seen centuries of change."
Story: "The clock tower in Oldham Square had always fascinated young Maya. Its face was cracked, and the bells no longer chimed, but it stood as a reminder of the village's rich history. Growing up, Maya loved listening to her grandfather's stories about how the tower once housed resistance fighters during the war. As an adult, Maya became a historian and decided to uncover the truth about these tales. She found old blueprints in the town archive showing hidden chambers within the tower. With the mayor's permission, she organized an excavation and discovered a dusty compartment containing letters and supplies from the resistance. The town celebrated her discovery, with a museum exhibit showcasing the tower's role in history. Maya's childhood wonder had turned into a significant contribution to her community. The tower was restored, and the bells rang once more, reminding everyone of the courage of their ancestors. It became a symbol of resilience and pride."
Sentence: "Sometimes, waiting for the perfect moment means missing the right one."
Story: "Mark had been eyeing the ring for months. He had rehearsed his proposal to Jane countless times in his head but kept waiting for a special occasion. On their anniversary, he carried the ring in his pocket but decided the moment wasn't romantic enough. On a weekend getaway to the mountains, he thought it was too cold for Jane to enjoy it. A few weeks later, Jane excitedly told him she had accepted a job offer overseas. Mark's heart sank; he realized his hesitation had cost him the chance to share his feelings. He rushed to her house that night, ring in hand, ready to confess. Jane was surprised but smiled warmly, telling him it was never too late to say how you feel. They spent the next few days planning their long-distance relationship. Mark learned that perfect moments are created, not waited for, and vowed never to let fear hold him back again."
Now, convert the following story into a concise sentence.
Story Input: Here is the story: {story}

Figure 2