KnowledgePrompts: Exploring the Abilities of Large Language Models to Solve Proportional Analogies via Knowledge-Enhanced Prompting

Thilini Wijesiriwardene[1], Ruwan Wickramarachchi[1], Sreeram Vennam[2], Vinija Jain[4]*, Aman Chadha[3]†, Amitava Das[1], Ponnurangam Kumaraguru[2], Amit Sheth[1]
[1]AI Institute, University of South Carolina, USA, [2]IIIT Hyderabad, India
[3]Amazon GenAI, USA, [4]Meta AI, USA
Correspondence: thilini@sc.edu

Abstract

Making analogies is fundamental to cognition. Proportional analogies, which consist of four terms, are often used to assess linguistic and cognitive abilities. For instance, completing analogies like "Oxygen is to Gas as <blank> is to <blank>" requires identifying the semantic relationship (e.g., "type of") between the first pair of terms ("Oxygen" and "Gas") and finding a second pair that shares the same relationship (e.g., "Aluminum" and "Metal"). In this work, we introduce a 15K Multiple-Choice Question Answering (MCQA) dataset for proportional analogy completion and evaluate the performance of contemporary Large Language Models (LLMs) in various knowledge-enhanced prompt settings. Specifically, we augment prompts with three types of knowledge: exemplar, structured, and targeted. Our results show that despite extensive training data, solving proportional analogies remains challenging for current LLMs, with the best model achieving an accuracy of 55%. Notably, we find that providing targeted knowledge can better assist models in completing proportional analogies compared to providing exemplars or collections of structured knowledge. Our code and data are available at: https://github.com/Thiliniiw/KnowledgePrompts/

1 Introduction

The ability to form analogies enables humans to transfer knowledge from one domain to another, making it a core component of human cognition (Hofstadter, 2001; Holyoak et al., 2001; Minsky, 1988). Specifically, in analogy-making, the emphasis is on the relations among objects, as it is the system of relations that is compared across domains rather than the specific objects and their attributes (Gentner, 1983). Researchers have identified several types of analogies within the domain of NLP, such as proportional analogies (analogies among word/term pairs) (Brown, 1989; Chen et al., 2022; Ushio et al., 2021; Szymanski, 2017; Drozd et al., 2016), sentence-analogies (Jiayang et al., 2023; Afantenos et al., 2021; Zhu and de Melo, 2020; Wang and Lepage, 2020) and analogies of longer text (Sultan and Shahaf, 2022; Sultan et al., 2024). Proportional analogies, which is the focus of this paper, are presented in the form *A:B::C:D*, meaning *A is to B as C is to D*. These analogies involve four terms, where the relationship between the first pair of terms (A and B) is similar to the relationship between the second pair of terms (C and D).

Generative Artificial Intelligence (GenAI) models, particularly those recognized for their capacity to generate high-quality textual outputs[1], have emerged as a focal point of research in contemporary Natural Language Processing. The capabilities of these models are typically evaluated through a range of tasks, including question answering (Arora et al., 2022; Kasai et al., 2023), reasoning (Zhang et al., 2024), paraphrasing (Witteveen and Andrews, 2019), sentiment analysis (Kheiri and Karimi, 2023) and, more recently, analogical reasoning (Bhavya et al., 2024; Wijesiriwardene et al., 2023). Notably, Wijesiriwardene et al. (2023) have demonstrated that SAT-style[2] Proportional analogies pose significant challenges for LLMs, particularly when solved using intrinsic distance-based similarity measures. Conversely, Webb et al. (2023) have shown that GPT-3 can surpass human performance in solving proportional analogies, though these findings were based on a dataset with limited size (774 data points) and a narrow range of dis-

*Work does not relate to position at Meta.
†Work does not relate to position at Amazon.

[1]In this work, Generative AI models refer to Large Language Models (LLMs) capable of producing high-quality textual content. Therefore, we use the term "GenAI Models" and LLMs interchangeably.

[2]SAT is a US college admission test where proportional analogies were used to assess linguistic and cognitive abilities of examinees.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3979–3996
January 19–24, 2025.

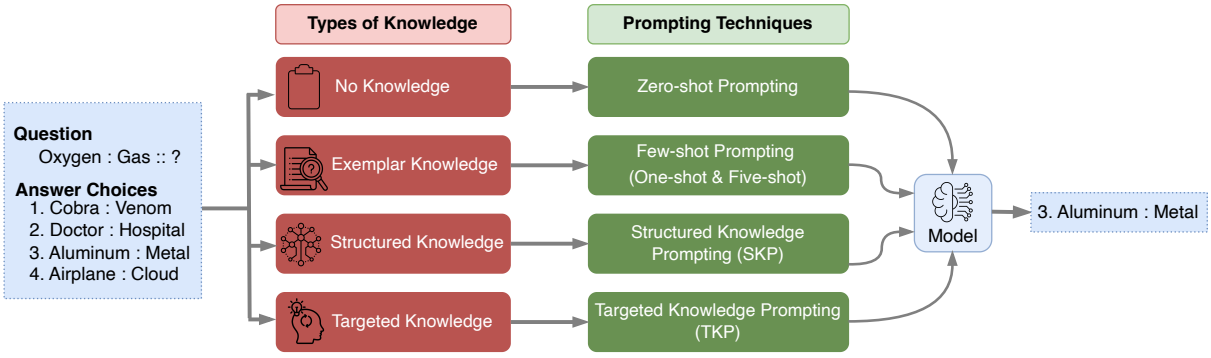

Figure 1: **Knowledge-enhanced Prompting.** An illustration of our knowledge-enhanced prompting approach with types of knowledge and prompting techniques. The question consists of two terms ("Oxygen" and "Gas"), and answer choices consist of term pairs that are analogous to the question term pair. Each model is queried using the prompting techniques illustrated.

tinct semantic relations among term pairs (seven semantic relation types). Motivated by the need to broaden the scope of research, we scale up the evaluation by assessing a diverse set of GenAI models on a larger, more comprehensive proportional analogy dataset. Additionally, we employ various prompting techniques enhanced with multiple types of knowledge to understand model capabilities in completing proportional analogies.

Our primary contribution lies in conducting a comprehensive evaluation of nine GenAI models, specifically assessing their performance in solving proportional analogies presented in a multiple-choice format. Considering the limitations of existing proportional analogy datasets, which typically comprise fewer than a thousand data points and a restricted range of relation types, we present a substantially larger dataset. Our dataset contains 15K proportional analogies with 236 distinct relation types. We evaluate the nine GenAI models on the 15K dataset using four distinct prompting techniques: (i) **Zero-shot Prompting**, where no additional knowledge is incorporated into the prompt, (ii) **Few-shot Prompting**, where exemplar knowledge in the form of examples from the dataset is included in the prompt, (iii) **Structured Knowledge Prompting (SKP)**, where the prompt is augmented with structured knowledge in the forms of lexical, commonsense, and world knowledge drawn from WordNet (McCrae et al., 2019), ConceptNet (Speer et al., 2017), and Wikidata (Vrandečić and Krötzsch, 2014) respectively and (iv) **Targeted Knowledge Prompting (TKP)**, which integrates targeted knowledge in the form of specific semantic relationships necessary for solving proportional analogies, along with the cognitive process behind such reasoning. To the best of our knowledge, this study is the first to explore knowledge-enhanced prompting strategies for solving proportional analogies.

Our findings indicate that completing proportional analogies is highly challenging for current LLMs and incorporating targeted knowledge significantly enhances model performance, with the best-performing model showing an improvement of approximately +21% compared to prompts without any knowledge, and around +45% relative to prompts enhanced with structured knowledge. The underperformance of SKP relative to Zero-shot Prompting suggests that the mere inclusion of relevant knowledge may not always improve model performance.

2 Related Work

In this section, we introduce related literature on the main topics of our paper: proportional analogies and LLMs, prompting techniques, and knowledge-enhancement in LLM prompting.

2.1 Proportional Analogies and LLMs

One of the earliest methods for solving proportional analogies was Latent Relational Analysis (LRA), introduced by Turney (2005). LRA determines analogy by measuring the similarity in semantic relationships shared between word pairs, considering them analogous if they exhibit a high degree of relational similarity. With the advent of neural networks, vector difference-based methods (Vylomova et al., 2016; Allen and Hospedales, 2019; Mikolov et al., 2013) were used to address proportional analogies. As LLMs based on the Transformer architecture (Vaswani et al., 2017a)

gained prominence, researchers began investigating the potential of LLMs, particularly Generative Artificial Intelligence (GenAI) models, for solving proportional analogies (Brown, 2020; Ushio et al., 2021; Webb et al., 2023). Specifically, Webb et al. (2023) demonstrated strong performance using a single model (GPT-3) on four relatively small proportional analogy datasets. Our study extends this work by scaling up the evaluation to a substantially larger dataset and by assessing nine contemporary GenAI models across six distinct prompting approaches. Additionally, we introduce a novel exploration of the impact of incorporating various types of knowledge when evaluating GenAI models on proportional analogies.

2.2 Prompting and Knowledge-enhanced Prompting

GenAI models are built on LLMs that are trained on extensive datasets and optimized for various tasks, including question-answering. This training implies that these models encapsulate the knowledge in the data, allowing them to effectively answer natural language queries (Roberts et al., 2020; Zhu and Li, 2023). Prompting involves transforming an input query into a structured natural language statement (prompt) and presenting it to the model, which then guides the output generation process of the model. (Schulhoff et al., 2024; Hadi et al., 2023; Liu et al., 2023). Generating outputs through prompting requires only forward passes during inference time, without any weight updates. Prompts can be created either manually (Wei et al., 2022; Schulhoff et al., 2024) or automatically (Ye et al., 2023; Reynolds and McDonell, 2021; Deng et al., 2022); in this work, we employ the more intuitive manual approach.

Prompts can be categorized based on the context they provide. Zero-shot prompts (Brown, 2020) contain only instructions related to solving a specific task, whereas Few-shot prompts (Brown, 2020) include both the instructions and one or more examples. Providing examples when querying models is a paradigm broadly known as In-context Learning (ICL) (Brown, 2020). Chain-of-Thought (CoT) Prompting is designed to guide models through the reasoning process required to solve a task by presenting an exemplar that includes the question, reasoning path, and correct answer (Wei et al., 2022) or by just incorporating a thought-inducing phrase such as "Let's think step by step" (Kojima et al., 2022) (Zero-shot-CoT). Unlike conventional CoT prompting, which often includes an exemplar, our adaptation termed TKP does not provide an exemplar. Instead, it enhances the prompt with the targeted knowledge specific to solving proportional analogies. As a result, TKP is more akin to Zero-shot-CoT (Kojima et al., 2022) than to traditional CoT (Wei et al., 2022).

The enhancement of LLM performance through the integration of external knowledge, both unstructured and structured, has been extensively studied (Yu et al., 2022). Some approaches transform external knowledge from multiple documents into graph structures and utilize these graphs to enhance LLM querying (Wang et al., 2024). Additionally, some methods directly employ structured knowledge (Baek et al., 2023). Retrieval-augmented generation (RAG) has recently emerged as an umbrella term encompassing all these techniques, where user queries are enriched with content retrieved from external sources to enhance model performance (Lewis et al., 2020; Ding et al., 2024; Mialon et al., 2023; Schulhoff et al., 2024). In this work, we utilize multiple types of knowledge, including targeted and structured knowledge (from three sources), to assess the impact on LLM performance in solving proportional analogies. To the best of our knowledge, this is the first study to explore the capabilities of LLMs in solving proportional analogies using knowledge-enhancement approaches.

3 Approach

As illustrated in Figure 1, given a proportional analogy MCQ where the question consists of a single term pair (`e.g., "Oxygen" and "Gas"`), the GenAI model is required to provide the correct answer choice from five, four or three choices. **Zero-shot Prompting**, only include the MCQ and a simple instruction on how to produce the output without any knowledge enhancement added to the prompt. Next, we enhance the Zero-shot Prompt with exemplars of solved MCQs from the dataset. We consider this approach as enhancing the prompt with "exemplar knowledge" and refer to this prompting technique as **Few-shot Prompting**. We experiment with one exemplar (One-shot Prompting) and five exemplars (Five-shot Prompting). Then a combination of lexical, commonsense, and world knowledge from structured sources—WordNet, ConceptNet, and Wikidata, respectively—is added to the Zero-shot

Prompts for knowledge enhancement, resulting in what we call **SKP**. Finally, the zero-shot prompt is enhanced with targeted knowledge and we identify this prompting technique as **TKP.** Targeted knowledge is composed of, the semantic relationship shared between the question term pair and the cognitive process behind solving the proportional analogy. We detail the prompting techniques in Section 3.3.

3.1 Dataset Creation

We introduce a 15K dataset of proportional analogies containing 5-way, 4-way and 3-way MCQs. Table 1 presents the dataset statistics along with examples from the dataset. We generate 14K questions out of the 15K based on the work by (Yuan et al., 2023). Yuan et al. (2023) introduced an automatically generated million-scale analogy knowledge base based on ConcepNet and Wikidata knowledge graphs. Yuan et al. (2023) acquire analogies of the same relations directly utilizing the concept pairs in the above-mentioned knowledge graphs. To acquire analogies of analogous relations (analogies consisting of two concept pairs with two relations that are analogous to each other) Yuan et al. (2023) utilize the in-context learning abilities of LLMs. We adopt this resource (specifically the analogies of same relations) to develop n-way (`n=[3, 4, 5]`) MCQs as follows. A single n-way MCQ consist of a pair of terms representing the *question* and `n` term pairs representing the *answer choices*, among which only one term pair is the correct answer. The semantic relationship between the term pair in the question is the same as the semantic relationship shared between the term pair which is the correct answer. The rest of the incorrect answer choices consist of term pairs with different semantic relationships among them.

Thousand data points out of the 15K are borrowed from work by Ushio et al. (2021); Turney and Littman (2003); Boteanu and Chernova (2015)[3] and contain 5-way, 4-way and 3-way MCQs. We highlight that, compared to previous proportional analogy MCQ datasets used for research (Webb et al., 2023; Ushio et al., 2021), the current dataset provides a significant increase in question quantity (∼15-times) and diversity (with respect to the diversity of semantic relations between terms). Our dataset also includes the semantic relationship shared by the question term pair compared to other datasets that do not include this information (Ushio et al., 2021; Turney and Littman, 2003; Boteanu and Chernova, 2015). Our dataset contains 59 semantic relationship types with more than 10 instances each. The distribution of these relationships (focusing on the top 59 types) is depicted in Figure 2.

3.2 Model Details

GenAI models are designed to generate content that are often indistinguishable from human-produced output. Current state-of-the-art GenAI models are largely based on the Transformer architecture (Vaswani et al., 2017b). In this work we compare the following popular open-source and proprietary GenAI models for their ability to solve proportional word analogy MCQs by incorporating variety of knowledge: (i) Falcon, a causal decoder-only model (Almazrouei et al., 2023), (ii) FlanT5 (Longpre et al., 2023), a T5 (Raffel et al., 2020a) based model trained on the Flan collection of datasets, (iii) GPT2 (Radford et al., 2019a), the first series of models to popularize in-context instructions, (vi) Mistral (Jiang et al., 2023), leveraging transformers architecture (Vaswani et al., 2017b) with several new introductions such as sliding window attention and pre-fill chunking, (v) Orca (Mukherjee et al., 2023), based on LLaMA model family (Touvron et al., 2023) and fine-tuned on complex explanation traces obtained from GPT-4 (Achiam et al., 2023), (vi) Zephyr (Tunstall et al., 2023), a fine-tuned version of Mistral trained on public datasets and optimized with knowledge distillation techniques. (vii) CodeT5 (Wang et al., 2021c), a unified pre-trained encoder-decoder transformer model leveraging code semantics and finally (viii) CodeParrot (Jain, 2023), a model based on GPT-2 and trained to generate python code (ix) GPT-3.5-Turbo [4]. Further details of the models used are presented in Appendix A.

3.3 Prompting Techniques

Currently, the most popular approach to Multiple Choice Question Answering (MCQA) is via cloze-style prompting (Brown, 2020; Robinson et al., 2023) where each answer choice is concatenated to the question separately and scored independently

[3]Unlike the 14K MCQs created based on AnalogyKB, these 1K data points do not provide the semantic relationship shared between the question term pair explicitly, therefore we employ two NLP researchers to discuss and manually identify the shared semantic relationship.

[4]https://platform.openai.com/docs/models/gpt-3-5-turbo

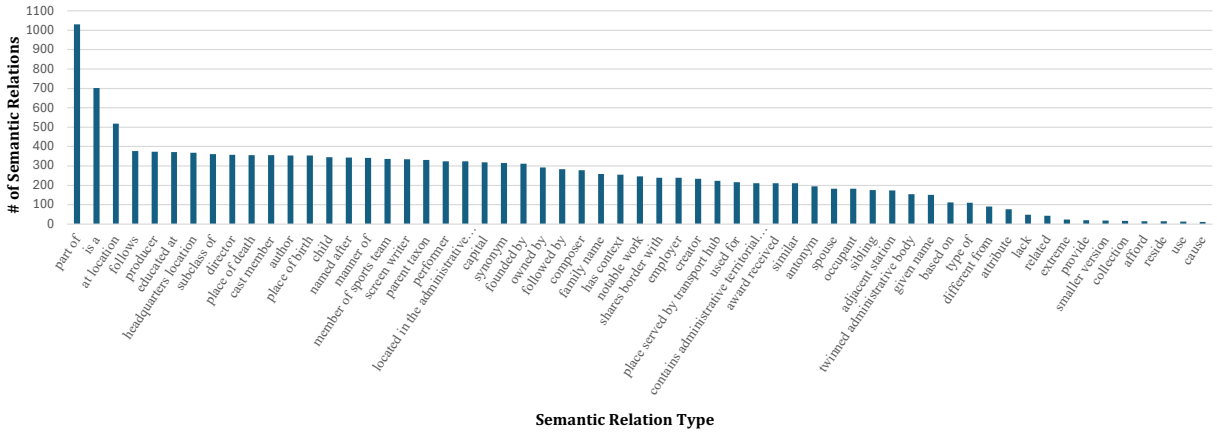

Figure 2: **Distribution of Semantic relations.** The distribution of the top 59 semantic relations (these are the frequencies of semantic relations between the question word pair)

	Questions			Relations	
Question Type (MCQ)	5-way	4-way	3-way	Top 5 Relation Types	# Data Points
Example	***Question***: Tenable: Indefensible ***Choices***: (1) Unique : Unprecedented (2) Dire : Pressing (3) Bleak : Desolate (4) Theoretical : Concrete (5) Recondite : Scholarly	***Question***: Haiku: Poem ***Choices***: (1) Song : Musician (2) Novel : Book (3) Artist : Painting (4) Page : Typeface	***Question***:Ancient: Old ***Choices***: (1) Crazy : Unhealthy (2) Delicious : Tasty (3) Smart : Intelligent	part of is a at location follows producer	1030 702 518 376 374
Amount	14386	610	4	**Total # relation types**	**236**

Table 1: **Dataset statistics**. The dataset consist of 15K MCQs that share 236 semantic relation types among them.

by the language model (LM). This style of prompting is problematic since it prevents the LM from comparing and contrasting all available options simultaneously. Additionally, it is computationally expensive, as it requires multiple forward passes through the LM to identify the correct answer (Robinson et al., 2023). To address these limitations, we adopt the prompt phrasing introduced by Robinson et al. (2023) with task-specific modifications. Specifically, the question and its symbol-enumerated candidate answers are provided to the model as a single prompt. Robinson et al. (2023) do not include specific instructions in the prompt for the model to output only the choice symbol. But we observe that adding such specific instructions reduce the model hallucinations. Therefore we use specific, non-ambiguous language to instruct the model to only output the relevant choice symbol. The prompting techniques are detailed below (See example prompts in appendix D).

3.3.1 Zero-shot Prompting

In Zero-shot Prompting, the question, all multiple choice answers and the instructions are provided in natural language (no knowledge is provided).

3.3.2 Few-shot Prompting

We demonstrate the task to the model by providing several exemplars in the form of question, answer choices and the correct answer choice. Then the actual question and answer choices are provided requiring the model to choose the correct answer choice. We employ one-shot and five-shot prompting under the few-shot prompting strategy where one example and five examples are provided respectively. We select these quantities of exemplars to strike a balance between the models' maximum accepted context length and the computational resources required. To obtain the exemplars, we employ a semantic similarity based filtering mechanism as follows. We encode each proportional analogy MCQ in the dataset using a SOTA sentence encoding transformer model[5], and identify the most semantically similar single example/ five examples based on Cosine similarity.

3.3.3 Structured Knowledge Prompting (SKP)

We retrieve knowledge from structured sources, filter it, and then integrate the resulting refined knowledge into the prompts. We detail this process

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

in the subsequent sections.

Knowledge Retrieval. We leverage the following widely-used large knowledge sources to obtain three types of knowledge: (i) Wikidata (Vrandečić and Krötzsch, 2014), which provides world knowledge in the form of explicit information about specific instances, encompassing billions of nodes and edges (Wang et al., 2021a); (ii) ConceptNet (Speer et al., 2017), a general-domain commonsense knowledge graph with 799,273 nodes and 2,487,810 edges; and (iii) WordNet (McCrae et al., 2019), a lexical database for the English language containing 161,338 words, 120,135 synsets, and 415,905 semantic relations.

We retrieve knowledge from above sources as follows. Since analogies focus on relations oppose to entities or entity attributes (Gentner, 1983), when retrieving knowledge from knowledge sources we focus on path finding approaches oppose to subgraph extraction approaches. To extract both world and commonsense knowledge, we utilize the path-finding approach by Lin et al. (2019) that identifies connections between each term pair (in both the question and answer choices). Specifically, we extract paths of length k[6] from ConceptNet and Wikidata. When retrieving lexical knowledge from WordNet, we extract the shortest path between term pairs.

Knowledge Filtering. For each term pair in the question and answer choices, multiple knowledge paths may be retrieved. To ensure the prompts stay within the maximum context length limit of the evaluated language models, we filter the retrieved paths and retain a single path for Wikidata and ConceptNet (See Figure 3). Filtering is not performed on WordNet since a single path (shortest) is always retrieved.

The filtering mechanisms we employ are as follows: (i) **Random Filtering**, where one path is randomly selected from the list of available paths; and (ii) **Semantic Filtering**, which selects the path most semantically similar to the term pairs. The term pairs (in question and answer choices) are formatted to "term pair sentences" in the following form <TERM_1> IS SEMANTICALLY RELATED TO <TERM_2> and returned paths are also formatted to "path sentences" in the form of [<NODE1_NAME> <RELATION1_NAME> <NODE2_NAME>, <NODE2_NAME> <RELATION2_NAME> <NODE3_NAME>, ...]. Both term pair sentences and path sentences are then encoded using a SOTA sentence encoding transformer model[7] and the path sentence with the highest cosine similarity to term pair sentence is filtered as relevant knowledge and referred to as *knowledge paths*[8].

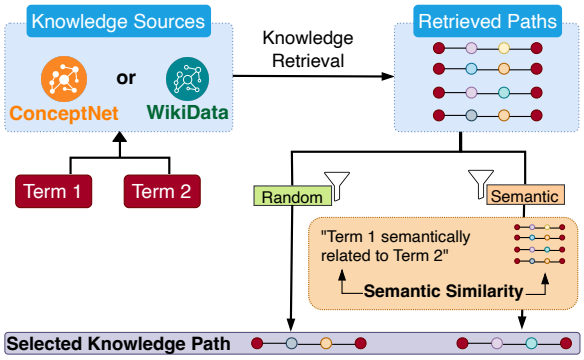

Figure 3: **An illustration of the knowledge filtering approach.** "Random" indicates Random Filtering and "Semantic" indicates Semantic Filtering.

Generating Prompt. The filtered knowledge paths are appended to the zero-shot prompt after the question and the answer choices to create the SKP and the model is instructed to use the knowledge if necessary. Based on the knowledge filtering mechanism SKP can be referred to as SKP[random] or SKP[semantic].

3.3.4 Targeted Knowledge Prompting (TKP)

When solving proportional analogies, humans typically examine the question term pair, identify the semantic relationship between the two terms, and select the answer pair that shares the same or a similar relationship. Inspired by this cognitive process, we modify the traditional Chain-of-Thought (CoT) prompting technique (Wei et al., 2022) to provide the model with "targeted knowledge" in the form of (i) semantic relationship shared by the question term pair (ii) cognitive process used by humans when evaluating such analogies, via the prompt.

[6] k is set to 2 for Wikidata and 3 for ConceptNet, as longer paths tend to introduce excessive noise and reduce efficiency.

[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[8] specific format of Wikidata knowledge paths is [<node1_name> <relation1_name> <node2_name>, <node2_name> <relation2_name> <node3_name>] and ConceptNet knowledge path is [<node1_name> <relation1_name> <node2_name>, <node2_name> <relation2_name> <node3_name>, <node3_name> <relation3_name> <node4_name>]

4 Experimental Setting

We have conducted a comprehensive set of experiments across nine GenAI models over six prompt variants on a 15K dataset, totalling to 54 ($9X6$) experiments. The implementation details are included in Appendix B

5 Results and Discussion

Proportional analogy multiple-choice questions (MCQs) are presented to each GenAI model using the previously described prompts. The model's response is extracted from the generated output, and accuracy is measured using Exact Match Accuracy (EMA) (Rajpurkar et al., 2016). While more flexible evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are commonly used to assess GenAI-generated outputs, we employ EMA because MCQs are inherently evaluated in a binary manner, where partial correctness is not rewarded. We report EMA as a percentage for each model and prompt variant. The results are presented in Table 2.

5.1 Model Performance and Prompting Techniques

The highest overall performance was attained by GPT-3.5-Turbo, achieving an EMA of 55.25%. This result underscores the challenge that proportional analogies pose for current state-of-the-art GenAI models. This accuracy was obtained through Targeted Knowledge Prompting where the prompt was enhanced with targeted knowledge (See Figure 5). Interestingly, the same model, when enhanced with structured knowledge, underperformed with an accuracy of 38% (EMA for SKP[random] is 38.29% and SKP[semantic] is 38.79%), compared to Zero-shot prompting (EMA 45.7%). This suggests that simply adding knowledge, even from diverse sources, may not be beneficial for cognitively demanding tasks such as proportional analogy completion. Out of the nine models four (Falcon, Flan, Mistral and GPT-3.5-Turbo) performs the best when prompted with Targeted Knowledge Prompts and two (GPT2 and Orca) performs the best with Zero-shot prompts with no knowledge enhancement. CodeT5 performs the best with one-shot prompts and Zephyr and CodeParrot performs the best with five-shot prompts. We also observe that models trained specifically on code generation such as CodeT5 and CodeParrot (specially CodeParrot) perform at the lower end of the spectrum despite the demonstrated abilities of them to perform well on other MCQ datasets Robinson et al. (2023). We believe this is due to the challenging nature of the proportional analogy completion task.

5.2 Role of Structured Knowledge in Model Performance

Although enhancing prompts with structured knowledge does not consistently improve model performance compared to other prompting techniques, SKP[semantic] leads to slight increases in EMA values (ranging from 0.01% to 1.32%) compared to SKP[random], across all models except GPT-2 and Mistral (see Table 2). We identified a subset of MCQs (19.96%) where all three types of knowledge were available and conducted additional experiments to evaluate the individual contribution of each knowledge type to EMA (we employed SKP[semantic] prompting). Our results show (see Figure 4, for complete results, see table 4 in Appendix C) that incorporating each of the three knowledge types separately into prompts leads to very similar EMA values (when averaged across all nine models). Specifically, prompts enhanced only with Wikidata knowledge resulted in an average EMA of 14.57%, while using only WordNet or only ConceptNet yielded average EMAs of 14.41% and 14.34%, respectively.

We also observed that incorporating all three types of knowledge simultaneously into the prompts, compared to using them individually, produced varying results. For example, Falcon, CodeT5 and GPT-3.5-Turbo perform marginally better when a single knowledge type is incorporated into the prompt, compared to including all three knowledge types simultaneously (see Figure 4). Providing FlanT5 with a single knowledge type compared to all three knowledge types contributes to significant increases of percentage points in EMA (WordNet +18.51 , ConceptNet +15.96 and WikiData +7.44). In contrast, GPT-2, Mistral, and Orca perform better when all knowledge types are integrated into the prompt. Notably, Orca demonstrates an average EMA increase of +11.14 percentage points compared to using only a single knowledge source.

5.3 Exemplar Quantity vs. Model Performance

Brown (2020) demonstrated that the accuracy of large language models improves with an increase

Model Name	Zero-shot Prompting	Few-shot Prompting		Structured Knowledge Prompting		Targeted Knowledge Prompting
		One-shot	Five-shot	Random	Semantic	
Falcon	24.17	23.21	22.61	24.75	25.01	**25.4**
FlanT5	36.47	40.09	38.07	14.43	14.62	**44.26**
GPT2	**22.65**	22.49	7.19	6.29	6.17	21.64
Mistral	26.59	26.22	27.34	24.58	24.42	**27.37**
Orca	**24.54**	23.28	14.11	18.48	18.81	24.2
Zephyr	29.46	34.05	**35.87**	16.13	17.22	15.83
CodeT5	20.64	**24.33**	0	16.15	17.47	21.64
CodeParrot	0	10.11	**12.6**	0	0.01	2.09
GPT-3.5-Turbo	45.7	31.79	41.21	38.29	38.79	**55.25**

Table 2: **MCQ Performance of models.** Performance is reported in EMA percentage. Best performance of each model is indicated in **bold** and the second best performance is indicated by underline.

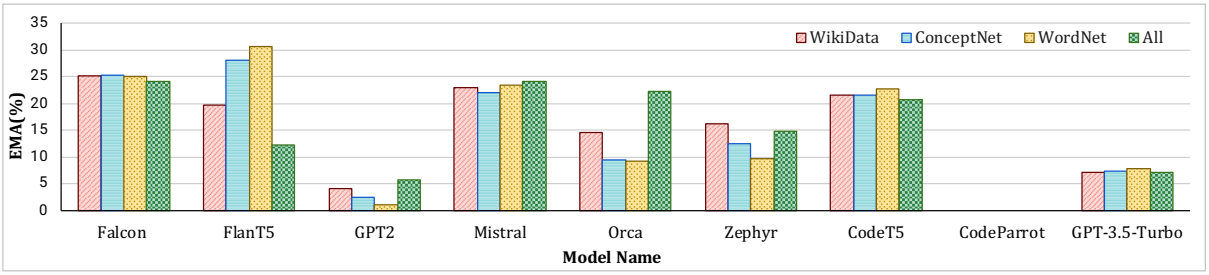

Figure 4: **Perfromance with structured knowledge.** Performance of each model when Structured Knowledge Prompting with semantic filtering (SKP[semantic]) is used. **All** indicates the prompt is enhanced with all three types of knowledge (Wikidata, ConceptNet and WordNet). EMA values are reported on 20% of the 15K dataset where all three knowledge types available.

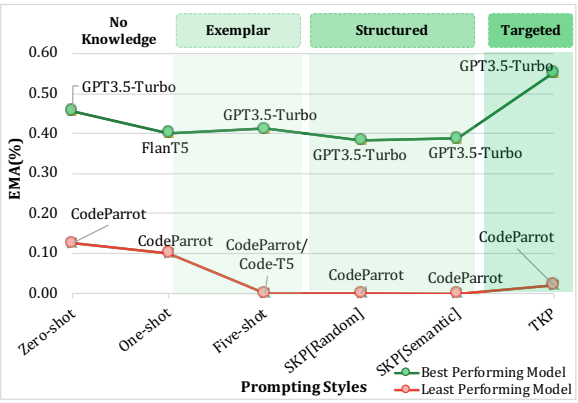

Figure 5: **Best and least performing models** for each prompting technique.

in the number of exemplars. However, Liu et al. (2022) found that the benefits diminish beyond 20 exemplars in certain cases. Similarly, in our study, increasing exemplars from one to five decreases EMA in six out of nine models (see Table 2), leading us to limit exemplars to a maximum of five.

5.4 Cost of Knowledge Acquisition vs. Model Performance

In this study, we utilize three types of knowledge to enhance prompts: exemplar knowledge, structured knowledge, and targeted knowledge. Among these, exemplar knowledge has the least acquisition cost since it is readily available from the dataset itself requiring no additional resources. Structured knowledge, on the other hand, is more expensive to acquire because it necessitates accessing external knowledge bases or graphs and filtering knowledge, which incurs computational overhead. Targeted knowledge is the costliest to acquire, as it involves identifying the specific semantic relationship between the question term pairs. This semantic relationship is not always readily available, requiring human annotation (for instance, in our dataset of 15K data points, 1K data points lacked this semantic information, necessitating human annotation).

As shown in Table 2, targeted knowledge, being the most expensive to acquire, led to the best performance in four models (Falcon, FlanT5, Mistral and GPT-3.5-Turbo) including the peak performance (55% EMA) from GPT-3.5-Turbo. In contrast, structured knowledge, the second most costly, did not result in any model's best performance. Although exemplar knowledge is the least expensive, three models performed best with it (Zephyr and CodeParrot in Five-shot; CodeT5 in One-shot).

Semantic Relationship Type	EMA % of Large Language Model								
	Falcon	FlanT5	GPT2	Mistral	Orca	Zephyr	CodeT5	CodeParrot	GPT-3.5-Turbo
part of	23.50	25.05	20.19	24.95	23.98	9.03	13.88	0.00	27.18
is a	25.78	29.20	20.09	25.50	22.08	6.84	19.37	0.00	31.34
at location	29.34	31.85	21.43	27.80	19.31	11.20	18.73	0.39	32.24
follows	22.07	28.46	15.43	22.34	23.94	3.19	14.36	0.27	35.11
producer	25.94	37.97	22.19	24.87	23.80	7.22	13.90	0.00	53.21
Avg. performance across the above relations	*25.33*	*30.51*	*19.87*	*25.09*	*22.62*	*7.50*	*16.05*	*0.13*	***35.82***

Table 3: **Performance of each LLM across semantic relations.** We report the performance of each LLM on the top five semantic relations in the dataset. For each LLM, the relation with the highest performance is highlighted in green, while the relation with the lowest performance is highlighted in orange. Additionally, the average performance across all five relation types is calculated and highlighted in grey.

5.5 Diversity of Semantic Relationships vs. Model Performance

As elaborated in Section 3.1, our dataset encompasses 236 unique semantic relation types, with the frequencies of the top five relations detailed in Table 1. To further elucidate the performance of each LLM, we assess their results on these top five semantic relations using targeted knowledge prompts (refer to Table 3). Consistent with prior findings, GPT-3.5-Turbo achieves the highest average performance across the top five relations, followed by FlanT5. For both models, the MCQs involving the "part of" relation pose the greatest challenge, whereas the "producer" relation is the easiest to solve. Similarly, across all nine LLMs, the "part of" and "follows" relations emerge as the most difficult, while the "at location" relation proves to be the easiest.

6 Conclusion and Future Work

We evaluate nine LLMs on a 15K MCQ dataset to assess their ability to solve proportional analogies using various knowledge-enhanced prompting techniques. Our experiments reveal that LLMs perform best when targeted knowledge is integrated into prompts, outperforming exemplar and structured knowledge.

While several of the models used are instruction-finetuned versions of their base models, they are not specifically finetuned for proportional analogy completion, leaving room for improvement. Additionally, our study focuses on manual prompting techniques, which are brittle; exploring automatic prompting approaches could yield more robust results.

7 Limitations

In SKP, knowledge paths may occasionally provide the exact semantic relationship between question term pairs, defined as targeted knowledge. These instances can be classified as SKP with targeted knowledge, but we do not currently verify or adjust for them. Also, in this work, we used manual prompt creation, where slight variations can significantly affect model outputs (Zhao et al., 2021). However, we did not address this variability by testing multiple prompt templates for each prompting technique. Acquiring targeted knowledge is resource-intensive due to the need for manual annotations. Scaling this process is impractical, highlighting the need for automated targeted knowledge acquisition techniques. Not all data points in the dataset include knowledge from ConceptNet, Wikidata, and WordNet due to the incompleteness of these graphs, highlighting the broader challenge of knowledge graph completion.

Acknowledgments

We thank anonymous reviewers for their constructive feedback and Anirudh Govil for his valuable input. This work was supported by NSF grant #2335967: EAGER: Knowledge-guided neurosymbolic AI with guardrails for safe virtual health assistants. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organization.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stergos Afantenos, Tarek Kunze, Suryani Lim, Henri Prade, and Gilles Richard. 2021. Analogies between sentences: Theoretical aspects-preliminary experiments. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21–24, 2021, Proceedings 16*, pages 3–18. Springer.

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings. In *International Conference on Machine Learning*, pages 223–231. PMLR.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. https://github.com/nomic-ai/gpt4all.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Re. 2022. Ask me anything: A simple strategy for prompting language models. In *The Eleventh International Conference on Learning Representations*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.

Bhavya Bhavya, Shradha Sehgal, Jinjun Xiong, and ChengXiang Zhai. 2024. AnaDE1.0: A novel data set for benchmarking analogy detection and extraction. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1723–1737, St. Julian's, Malta. Association for Computational Linguistics.

Adrian Boteanu and Sonia Chernova. 2015. Solving and explaining analogy questions using semantic networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Tom B Brown. 2020. Language models are few-shot learners. *Annual Conference on Neural Information Processing Systems*.

William R Brown. 1989. Two traditions of analogy. *Informal Logic*, 11(3).

Jiangjie Chen, Rui Xu, Ziquan Fu, Wei Shi, Zhongqiao Li, Xinbo Zhang, Changzhi Sun, Lei Li, Yanghua Xiao, and Hao Zhou. 2022. E-kar: A benchmark for rationalizing natural language analogical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3941–3955.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.06211*.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 3519–3530.

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

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Douglas R Hofstadter. 2001. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

K Holyoak, Dedre Gentner, and B Kokinov. 2001. The place of analogy in cognition. *The analogical mind: Perspectives from cognitive science*, 119.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*.

Royal Jain. 2023. codeparrot (CodeParrot).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Cheng Jiayang, Lin Qiu, Tsz 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

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.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Hoi. 2022. CodeRL: Mastering code generation through pretrained models and deep reinforcement learning. In *Advances in Neural Information Processing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Pankaj Mathur. 2023. orca_mini_7b: An explain tuned openllama-7b model on custom wizardlm, alpaca, and dolly datasets. `https://github.com/pankajarm/wizardlm_alpaca_dolly_orca_open_llama_7b`, `https://https://huggingface.co/psmathur/wizardlm_alpaca_dolly_orca_open_llama_7b`.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. *Preprint*, arXiv:2210.12353.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Oren Sultan, Yonatan Bitton, Ron Yosef, and Dafna Shahaf. 2024. Parallelparc: A scalable pipeline for generating natural-language analogies. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5900–5924.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562.

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 448–453.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Peter D Turney. 2005. Measuring semantic similarity by latent relational analysis. *arXiv preprint cs/0508053*.

Peter D Turney and Michael L Littman. 2003. Combining independent modules in lexical multiple-choice problems. In *Recent Advances in Natural Language Processing III*, page 101–110.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, and Aidan N Gomez. 2017a. L. u. kaiser, and i. polosukhin,"attention is all you need,". *Advances in neural information processing systems*, 30:5998–6008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is

all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Chenhao Wang, Yubo Chen, Zhipeng Xue, Yang Zhou, and Jun Zhao. 2021a. Cognet: Bridging linguistic knowledge, world knowledge and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 16114–16116.

Liyan Wang and Yves Lepage. 2020. Vector-to-sequence models for sentence analogies. In *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 441–446. IEEE.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021c. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859*.

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.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada. Association for Computational Linguistics.

Sam Witteveen and Martin Andrews. 2019. Paraphrasing with large language models. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 215–220, Hong Kong. Association for Computational Linguistics.

Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. 2023. Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2023. Analogykb: Unlocking analogical reasoning of language models with a million-scale knowledge base. *arXiv preprint arXiv:2305.05994*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*.

A Model Details

Falcon (Almazrouei et al., 2023): The Falcon model used in this work is the Falcon-7B-Instruct model [9] which is a causal decode-only model, instruction finetuned on top of the base Falcon-7B. The fine-tuning dataset is made up of 250M tokens from various conversational datasets (Baize[10]), instruction datasets (GPT4All (Anand et al.,

[9] https://huggingface.co/tiiuae/falcon-7b-instruct

[10] https://github.com/project-baize/baize-chatbot/tree/main/data

2023), GPTeacher[11]) and common crawl data (RefinedWeb (Penedo et al., 2023))from the web. Falcon-7B tokenizer is used for tokenization. The architecture of Falcon is broadly adapted from GPT3 with changes in positional embeddings used, attention mechanisms used and decoder block architecture.

FlanT5 (Chung et al., 2022): We use the FlanT5-XXL version with 11B parameters. This version is based on a pretrained T5 (Raffel et al., 2020b) and instruction finetuned on a mixture of tasks. This model is finetuned specifically with Chain-of-Thought data.

GPT2 (Radford et al., 2019b): We use the XL version with 1.5 parameters. The model is pretrained with English language data (40 GB of text from the web) and causal language modeling objective. Interestingly the model is not trained on articles from Wikipedia.

Mistral (Jiang et al., 2023): This is a decoder only transformer model and we use the Mistral-7B-Instruct version with 7B parameters. This version is finetuned on publicly available instruction datasets. Mistral introduce Sliding Window Attention, Rolling Buffer Cache and Pre-fill Chunking in its architecture.

Orca (Mathur, 2023): We employ orca_mini_7b, a 7B parameter version of Orca, which is based on OpenLLaMA-7B. The model is trained on datasets with explanation tuning, where the response from the <query, response> pair is augmented with detailed responses from the base (teacher) model (Mukherjee et al., 2023). The explanation tuning datasets used are WizardLM[12], Alpaca dataset (Taori et al., 2023) and Dolly[13] and system prompts are used to elicit step-by-step explanations.

Zephyr[14]: We use the Zephyr-7B-alpha with 7B parameters finetuned from Mistral-7B-v0.1. The finetune datasets contain synthetic dialogues ranked by GPT-4 and a prompt completion dataset where completions are ranked by GPT-4.

CodeT5 (Le et al., 2022): The CodeT5 model wwe use is codet5-large model with 770M parameters. The model is trained on Masked Span Prediction objective on CodeSearchNet dataset (Husain et al., 2019)

CodeParrot (Jain, 2023): We use the 1.5B parameter CodeParrot model based on GPT-2. The model is trained to generate python code on a python files dataset from GitHub[15].

GPT-3.5-Turbo[16]: We use OpenAI API to access the model, gpt-3.5-turbo-0125.

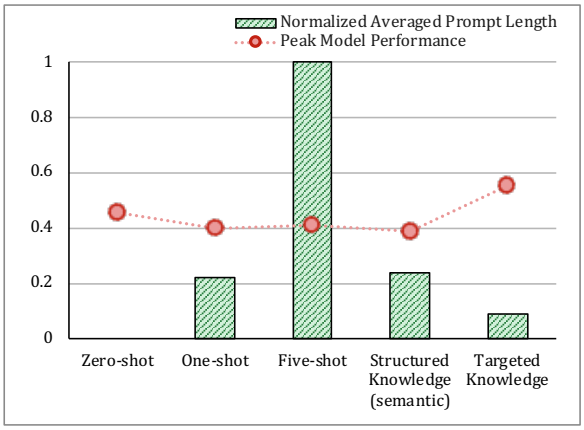

Figure 6: Prompt Lengths vs. Peak Performance.

B Implementation Details

We use API requests for GPT-3.5-Turbo and checkpoints from Hugging face[17] for open-source models. The models are evaluated with following hyper parameter settings, temperature = 0.1, top_p=0.1 and repetition_penalty=1.2 to elicit more concrete answers for the MCQs. We use Sentence Transformers[18] to identify semantically similar exemplars and to perform semantic knowledge filtering. We utilize Wikidata knowledge from (Wang et al., 2021b), ConceptNet knowledge from conceptnet5[19] and WordNet knowledge from Open English WordNet (2023)[20].

[11] https://github.com/teknium1/GPTeacher
[12] https://github.com/nlpxucan/WizardLM
[13] https://github.com/databrickslabs/dolly
[14] https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha
[15] https://huggingface.co/datasets/codeparrot/codeparrot-clean
[16] https://platform.openai.com/docs/models/gpt-3-5-turbo
[17] https://huggingface.co/models
[18] https://sbert.net/
[19] https://github.com/commonsense/conceptnet5/wiki/Downloads
[20] https://github.com/globalwordnet/english-wordnet?tab=readme-ov-file

Model Name	WD Knowledge Only	CN Knowledge Only	WN Knowledge Only	All Knowledge Available
Falcon	25.14	25.38	25.04	24.14
FlanT5	19.63	28.15	30.70	12.19
GPT2	4.11	2.40	1.00	5.84
Mistral	22.84	22.07	23.34	24.11
Orca	14.62	9.52	9.28	22.24
Zephyr	16.16	12.49	9.75	14.76
CodeT5	21.47	21.64	22.70	20.70
CodeParrot	0.00	0.00	0.00	0.00
GPT-3.5-Turbo	7.13	7.43	7.90	7.20

Table 4: **Performance of models based on provided knowledge types.** Performance values are reported in EMA percentage and calculated using 2995 (∼20%) data points that had all three knowledge types available.

C Performance and Additional Results

C.1 Model Performance vs. Prompt Length (PL)

We calculated the average prompt lengths across models for each prompting technique (PL for SKP is calculated by averaging SKP[random] and SKP[semantic]) (See Figure 6). According to (Liu et al., 2024), longer prompts (with important information placed in the middle) tend to negatively affect performance. Based on such literature, one might suggest that Zero-shot prompts yield better results in our study because they are short, but this is not the case. Despite being longer than Zero-shot prompts, a higher peak model performance is achieved by TKP.

D Prompts

Figues 7, 8, 9, 10 and 11 illustrates example prompts provided to models.

Zero-shot Prompt

```
Question: What is the analogical word pair to, "Lens" and "Glass" from the following choices.
Choices:
    1. "Well" and "Water"
    2. "Saw" and "Wood"
    3. "Sweater" and "Wool"
    4. "Fuel" and "Fire"
    5. "Ink" and "Paper"
The answer should only be 1 or 2 or 3 or 4 or 5?.
Answer:
```

Figure 7: Example of a Zero-shot prompt used on our dataset

One-shot Prompt

```
Look at the following example and answer the question below.

Example:
    Question: What is the analogical word pair to, "Cloth" and "Threads"      from the
following choices. Choices:
                1. "Gun" and "Bullets"
                2. "Guitar" and "Drums"
                3. "Chain" and "Links"
                4. "Star" and "Planets"
    Answer: 3

Question: What is the analogical word pair to, "Lens" and "Glass" from the following
choices. Choices:
    1. "Well" and "Water"
    2. "Saw" and "Wood"
    3. "Sweater" and "Wool"
    4. "Fuel" and "Fire"
    5. "Ink" and "Paper"
The answer should only be 1 or 2 or 3 or 4 or 5?.
Answer:
```

Figure 8: Example of a One-shot prompt used on our dataset

Five-shot Prompt

```
Look at the following examples and answer the question below.

Example 1:
    Question: What is the analogical word pair to, "Cloth" and "Threads" from the following
choices. Choices:
                1. "Gun" and "Bullets"
                2. "Guitar" and "Drums"
                3. "Chain" and "Links"
                4. "Star" and "Planets"
    Answer: 3
Example 2:
    Question: What is the analogical word pair to, "Drapery" and "Fabric" from the following
choices. Choices:
                1. "Fireplace" and "Wood"
                2. "Curtain" and "Stage"
                3. "Shutter" and "Light"
                4. "Sieve" and "Liquid"
                5. "Window" and "Glass"
    Answer: 5
Example 3:
    ……
Example 4:
    ……
Example 5:
    ……
Question: What is the analogical word pair to, "Lens" and "Glass" from the following choices. Choices:
    1. "Well" and "Water"
    2. "Saw" and "Wood"
    3. "Sweater" and "Wool"
    4. "Fuel" and "Fire"
    5. "Ink" and "Paper"
The answer should only be 1 or 2 or 3 or 4 or 5?.
Answer:
```

Figure 9: Example of a Five-shot prompt used on our dataset

Structured Knowledge Prompt

```
Question: What is the analogical word pair to, "Lens" and "Glass" from the following
choices. Choices:
    1. "Well" and "Water"
    2. "Saw" and "Wood"
    3. "Sweater" and "Wool"
    4. "Fuel" and "Fire"
    5. "Ink" and "Paper"
The answer should only be 1 or 2 or 3 or 4 or 5?.
Use following knowledge to find the correct answer choice.
Question Knowledge: glass related to device, device is a camera, camera is a lens; optical device is
a device, device is a instrumentality, instrumentality is a container, container is a glass
Answer choice 1 knowledge: well is a place, place is a water_route, water_route is a water
Answer choice 2 knowledge: wood made of house, house made of metal, metal made of saw; power tool is
a machine, machine is a device, device is a instrument, instrument is a wind, wind is a wood
Answer choice 3 knowledge: coat made of wool, wool related to warm, warm related to sweater; garment
is a habiliment, habiliment is a covering, covering is a artifact, artifact is a textile, textile is
a wool
Answer choice 4 knowledge: fuel is a substance, substance is a element, element is a fire
Answer choice 5 knowledge: ink at location sign, sign at location paper; ink is a change, change is a
cover, cover is a paper
Answer:
```

Figure 10: Example of a Structured Knowledge Prompt[semantic] used on our dataset

Targeted Knowledge Prompt

```
Question: What is the analogical word pair to, "Lens" and "Glass" from the following choices. Choices:
    1. "Well" and "Water"
    2. "Saw" and "Wood"
    3. "Sweater" and "Wool"
    4. "Fuel" and "Fire"
    5. "Ink" and "Paper"
The answer should only be 1 or 2 or 3 or 4 or 5?.
The implicit relation shared by "lens" and "glass" is "made of". The correct choice should have the same implicit
relation among the two words.
Answer:
```

Figure 11: Example of a Targeted Knowledge Prompt used on our dataset