From Prototypical to Relational: How LLMs Navigate Complex Analogies

Mayukh Das and **Wolf-Tilo Balke**
Institute for Information Systems, TU Braunschweig
Mühlenpfordtstraße 23, 38106 Braunschweig, Germany
{mayukh,balke}@ifis.cs.tu-bs.de

Abstract

We introduce a comprehensive benchmark to assess the analogical reasoning capabilities of large language models (LLMs) on complex analogy tasks that go beyond conventional formats with single correct answers. Unlike standard benchmarks that assume a singular ground truth, our framework presents a four-way multiple-choice analogy task in which all target options are semantically plausible. Leveraging concept pairs from Wikidata and AnalogyKB, we construct analogy instances enriched with multiple overlapping relational structures, where the relations are mined with RAG and ranked in salience through a GPT-4-assisted Max-Diff survey. To enable systematic evaluation, we propose three complementary semantic measures i.e. ranked relational overlap, context embedding similarity, and prototypicality; each grounded in established literature on analogical reasoning. Our experiments span a range of LLMs, evaluated under zero-shot and knowledge-enhanced prompting conditions. While models such as GPT-4 perform well on embedding-based and prototypicality-based measures, they consistently underperform when tasked with capturing fine-grained relational mappings. These results reveal that, despite their impressive surface-level semantic fluency, current LLMs exhibit notable limitations in structured relational reasoning.

1 Introduction

Analogical reasoning is the cognitive ability to recognize, map, and apply structural relationships between seemingly disparate concepts. It is widely regarded as a cornerstone of human intelligence, creativity, and abstraction (Gentner, 1983). By enabling the transfer of knowledge from a familiar domain (the source) to a novel one (the target), analogy not only facilitates comprehension but also supports reasoning, explanation, and flexible problem-solving across a wide range of contexts (Hofstadter, 2013). Beyond simple comparisons, analogical reasoning provides a powerful mechanism for generating hypotheses, fostering conceptual change, and driving scientific discovery. For instance, the classical analogy comparing an atom to a solar system—where electrons orbit the nucleus as planets orbit the sun, illustrates how analogical thinking scaffolds the understanding of abstract or counterintuitive scientific ideas by relating them to more familiar and intuitive phenomena. Such mappings, even if not scientifically precise, often serve as conceptual entry points that guide learners toward deeper theoretical insights. More broadly, analogical reasoning is central to language, metaphor, and creativity, underpinning the ability to extend prior knowledge to novel situations and to reframe problems in innovative ways.

Figure 1: Example of a complex analogy with multiple plausible targets, illustrating the need for fine-grained relational reasoning. LLMs does not lean towards relational structure based semantics

Humans can navigate complex systems through familiar relational templates (Hofstadter, 2013). Recent research in natural language processing (NLP) has posited that analogical reasoning similarly enables large language models (LLMs) to generalize beyond explicit training examples (Yasunaga et al., 2024). Accordingly, numerous

Proceedings of the 18th International Natural Language Generation Conference, pages 465–485
October 29–November 2, 2025.

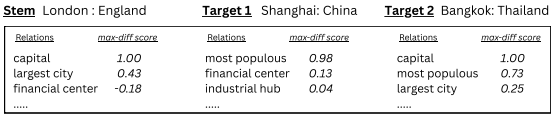


Figure 2: Concept pair has multiple ranked relations

studies have explored the analogical capacities of LLMs across tasks such as 4-way analogies (Ushio et al., 2021b), narrative analogy, story analogy, and relation-mining (Young et al., 2022; Zhou et al., 2024; Yuan et al., 2024). These findings often report that LLMs can solve analogy problems in zero-shot settings, exhibiting behaviors loosely aligned with human analogical inference (Webb et al., 2023; Kojima et al., 2023).

However, these capabilities remain fragile and inconsistent and are highly sensitive to prompt design, context, and task formulation. Crucially, prior work has largely focused on traditional analogy benchmarks like 4-term analogy or SAT-style tasks (Turney et al., 2003), which feature a single correct answer and relatively shallow relational mapping (only one relation of stem matches with the correct target). While recent studies have proposed new analogical datasets and tasks (Yuan et al., 2024; Sehgal et al., 2024), they often fall short in providing a sufficiently detailed semantic framework, which is essential for systematically guiding and evaluating task performance. This paper addresses a critical gap in the literature by investigating how generative large language models (LLMs) perform on analogy tasks with multiple plausible solutions, framed in an SAT-style format. We examine whether generative LLMs engage in relational reasoning, consistent with established cognitive theories of analogy (Christoph Lofi, 2013), or whether their responses reflect a reliance on prototypical, surface-level features. To this end, we make the following contributions:

1. **A Novel Benchmark for Complex Analogical Reasoning:** We construct a set of 1210 samples, four-way multiple-choice analogy dataset in which all target choices are viable analogical matches to the stem. A key feature of this dataset is that each concept pair has multiple relations that links the pair and these relations are mined by Retrieval Augmented Generation (Yu et al., 2025) and ranked by salience through a GPT-4-driven Max-Diff survey (Louviere et al., 2015). For example, the pair "London" and "England" may share multiple relations such as *largest_city*, *capital_of*, and *financial_center*. These relations can be ranked by importance (see Fig. 2), where *capital_of* holds higher salience than *largest_city*, which in turn ranks above *financial_center*.

2. **Ground Truth depends on specific Semantics:** Unlike prior benchmarks with rigid labels, we determine the most appropriate target using one of three proposed semantic measures—*ranked relational overlap*, *context embedding similarity*, and *prototypicality*; each grounded in prior cognitive and computational research. This allows us to investigate which measure best aligns with LLM predictions and to disambiguate the types of reasoning LLMs tend to rely on.

3. **Semantically Plausible Distractor Design:** We ensure that each target shares at least one relation with the stem, thereby increasing task difficulty and realism. Unlike conventional datasets that rely on strictly incorrect distractors, our approach makes all options semantically defensible and shifts the focus to precise semantic matching.

Taken together, our results highlight a critical limitation in current LLM architectures: their analogical reasoning capabilities are primarily grounded in shallow statistical or prototypical associations rather than deep relational alignment. Although prompting techniques like knowledge-enhanced prompting offer improvements, they do not fully address this limitation. Our findings suggest limitations in current claims regarding emergent analogical reasoning in LLMs (Webb et al., 2023; Yasunaga et al., 2024) and highlight the potential value of incorporating more explicit mechanisms for relational abstraction in future models. All data, annotations, and evaluation scripts are publicly available[1].

2 Related Work

The performance of language models on analogical reasoning has been a growing area of research, showing significant advancements over time.

Model Performance on Analogy Task: This work assesses model performance on analogy tasks

[1] https://github.com/Mayukhga83/Analogy-Task

using novel prompting, task descriptions, finetuning, and other techniques. Earlier research employs the word analogy task to assess the analogical reasoning abilities of language models (Mikolov et al., 2013a; Levy and Goldberg, 2014; Fournier et al., 2020; Ushio et al., 2021b). Recent studies (Yasunaga et al., 2024) have introduced analogical prompting, where language models are guided to generate relevant reasoning exemplars before solving a problem. Yung et al. (2022) developed prompts based on structured mapping theory and explored whether models can abduce structure while concept mapping. Zhou et al. (2024) introduced link-of-analogy prompting, which enables LLMs to process new situations by drawing analogies to known situations. Some studies posit that highly scaled models, such as GPT3, may achieve performance levels comparable to those of humans; however, this performance is often task-specific (Webb et al., 2023; Hu et al., 2023; Wijesiriwardene et al., 2023; Jiayang et al., 2023).

Analogy curation: Early studies primarily obtain analogy knowledge through the expertise of linguists (Adrian Boteanu, 2015). Later studies consider exploiting relations in common sense Knowledge Graphs to curate analogies (Allen and Hospedales, 2019; Ulčar et al., 2020; Speer et al., 2008; Li et al., 2018; pen; Gladkova et al., 2016; Zhang et al., 2019; Ilievski et al., 2022). These studies are characterized by either suboptimal quality or high quality but with very limited sample sizes. To tackle such problems Yuan et al. (2024) curated a large-scale analogy knowledge base derived from existing knowledge graphs. Jurgens et al. (2012) tried to identify the degree of prototypicality for word pairs within a given relation class. Unlike previous studies, this research focuses solely on analogy examples with multiple plausible solutions. Furthermore, it is the first to systematically examine the correlation between LLM predictions on analogy tasks and semantic measures, providing new insights into their reasoning processes.

3 SAT Multiple Choice Analogy Task

A multiple-choice four-way analogy task (Turney et al., 2003; Ushio et al., 2021a,b) involves presenting an analogy problem consisting of a pair of related words or concepts as the stem, followed by several word pair options as target choices (see Figure 1). The goal is to find the target pair that best aligns with the stem, as solving analogy tasks requires matching relational structures across pairs. Turney et al. (2003) collected 374 multiple-choice questions from SAT exams where each question has one stem and five targets. In this section, we evaluate the complete dataset using decoder LLMs of varying scales, including the open-source LLaMA 2 (13B and 70B model) (Touvron et al., 2023) and GPT3.5, GPT4 (OpenAI, 2023) and compare it with previous benchmarks done by Petersen and van der Plas (2024) on encoder LLMs. Petersen and van der Plas (2024) in their benchmark used two variants of Bert, one where $cos(a_1 - a_2, b_1 - b_2)$ is minimized (Bert-aa) and one where $cos(a_1 - b_1, a_2 - b_2)$ is minimized (Bert-ab) for 4-way $a_1 : a_2 :: b_1 : b_2$ analogy task[2]. We benchmarked the decoder models under both zero-shot and one-shot settings to assess the extent of performance improvement across these configurations. For the one-shot setting, we designed prompts using techniques like Chain of Thought (CoT) (Wei et al., 2022b) and Automatic Chain of Thought (Auto-CoT) (Zhang et al., 2022). For details on the prompts, see appendix 9.1. The results are summarized in Figure 3. Our findings indicate that small-scale decoder LLMs achieve performance comparable to both human accuracy and encoder models. In contrast, large-scale decoder models, such as GPT4, surpass human-level performance by a substantial margin, highlighting a significant scaling advantage. Furthermore, incorporating a single in-context example in a Chain-of-Thought (CoT) framework yields only a marginal performance improvement in GPT models. These findings suggest that analogical reasoning in LLMs likely emerges implicitly as a byproduct of increased model scale and extensive data exposure, rather than through explicit programming, task-specific training, or the introduction of in-context examples (Wei et al., 2022a; Webb et al., 2023). Building on this insight, this study systematically examines model performance on more challenging analogy tasks, providing a deeper analysis of their reasoning abilities and limitations.

4 Task Formalization

Christoph Lofi (2013) provided a foundational framework for analogy tasks. Expanding on this, we investigate a four-way multiple-choice analogy

[2]Please note we have adapted the notation used by Petersen and van der Plas (2024) to align with the notation used throughout this paper

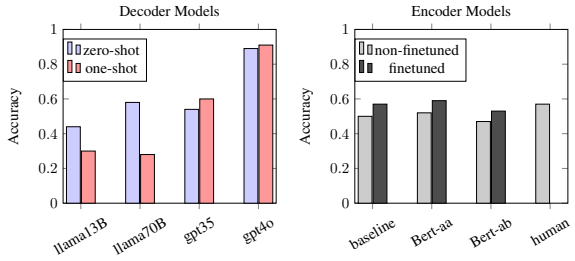


Figure 3: LLM accuracy on the SAT analogy test: Left plot (our decoder benchmark) vs. right plot (Petersen and van der Plas (2024)'s encoder). Baseline: FastText (Bojanowski et al., 2017) (non-finetuned) and BERT classifier (finetuned)

task. Each problem consists of a stem concept pair and four target concept pairs, with the goal of selecting the target pair that most closely aligns with the stem pair. However, in our method multiple target pairs may appear plausible, and the correct choice is determined based on predefined criteria, referred to as semantic measures.

4.1 Analogy Concept Pairs

Christoph Lofi (2013) defined concept analogy concept pairs as a subset of the power set of all concepts $A_{full} \subseteq P(C)$ where C is the set of all possible concepts. While they referred to these as analogons, we adopt the term analogy concept pairs for clarity and ease of interpretation. In case of designing analogy task, only a restricted subset of analogy concept pairs, which includes precisely two concepts, is utilized

$$A_{restricted} \subseteq C \times C \subseteq A_{full}$$

From this point, whenever we mention concept pair we mean elements of $A_{restricted}$. The conceptual operations necessary for our task definition, involves four steps.

4.2 Relevant Relations of each Concept Pair

Retrieve the set of relevant relationships of a concept pair. If $a \in a_1, a_2$ be a concept pairs. Let $r^o(a_1, a_2)$ be the set of all possible relationships between a_1 and a_2, then the set of relevant relationships between them is a space of restricted relation

$$r_a = r^o(a_1, a_2) \setminus \{\kappa\}$$

where κ is a set of elements that is removed from $r^o(a_1, a_2)$ as per some filtering criteria.

4.3 Relevant Relation Score

This evaluates the strength of a relation with respect to its parent concept pair, enabling the ranking of relevant relations according to their significance within the parent concept pair. If S is some scoring mechanism and r_a has n relations then

$$S(r_a) = \{(r_1^a, s_1^a), (r_2^a, s_2^a), (r_3^a, s_3^a)...(r_n^a, s_n^a)\}$$

where r_i^a is the i^{th} relation between concept pairs a_1 and a_2 and s_i^a is its score. The ranking aims to measure the depth of the relational importance between two concept pairs.

$$r_i^a = r_j^b \not\Rightarrow s_i^a = s_j^b$$

Because if the i^{th} relation between a_1 and a_2 is similar to the j^{th} relation between b_1 and b_2, that does not imply their respective importance is same for both the concept pair. The same relation may be more relevant to its parent concept pair than to another.

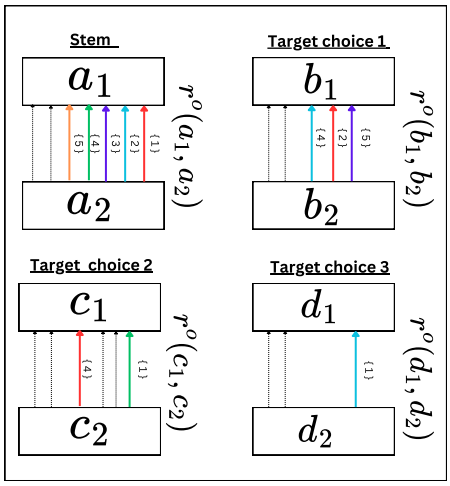


Figure 4: Stem and target pairs share many overlapping relations, distinguished by color codes, but the ranks of these relations vary between individual concept pairs.

4.4 Stem-Target Relation Overlap

Let $a \in a_1, a_2$ and $b \in b_1, b_2$ be two concept pairs, stem and target respectively. Then the set of overlapping relationships between the two pairs are

$$r_{ab} = r_a \cap r_b : r_{ab}, r_a, r_b \subseteq R$$

4.5 Semantic Measure

The purpose of the measure is to interpret the correct answer out of multiple plausible target analogons.

$$M(r_{ai}, S(r_a), S(r_i)) = c_i \forall i \in T$$

Where r_a is interpreted as relevant relations of the stem concept pair, T is the set of all possible target concept pairs and r_i is the relevant relations of the i_{th} target concept pair. Thus, the measure calculates a confidence score $c_i \in (0, 1)$ by evaluating the overlap between stem and target relations along with their scores, determining how strongly each target aligns as an analogy of the stem. The specific properties of the chosen measure determine how the relations are incorporated—for example, whether through mathematical operations on relation scores or by matching relational contexts.

5 Dataset Curation

5.1 Identify Semantically Rich Relations

To curate analogy examples with rich relational semantics, we begin by identifying concept pairs connected through relations that are likely to co-occur with additional meaningful links. We refer to these as *semantically rich relations*. Drawing from Wikidata (Vrandečić and Krötzsch, 2014), we selected 18 such relations based on their empirical tendency to overlap with diverse relational types. In Wikidata, relations are encoded as structured property triples—for example, {Q90: Paris, P36: capital of, Q142: France}, where *capital of* is the relation linking *Paris* and *France*. Crucially, this pair is also associated with other semantically salient relations such as *located in* (P131), enriching the conceptual linkage. By contrast, pairs such as {Q1048: Cleopatra, P106: occupation, Q82955: politician} involve only a single, narrowly defined relation and thus lack comparable semantic depth.

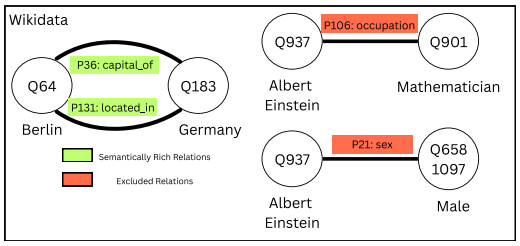


Figure 5: Illustrating semantically rich relations

These become our root relations, using which we gather concept pairs. Our selected relation types include categories such as *operating system*, *part of*, *city-country*, and *shares border with*, among others. This yields a total of 20 base relations used in our benchmark. Appendix 9.2 lists these relations.

5.2 Gather Concept Pairs

After identifying semantically rich relations, we first collected multiple concept pairs for each of these root relations. Then after we had all the concept pairs, we mined all other possible relations for each concept pair using RAG (section 5.3).

5.2.1 Capital and Location Relations

We extracted concept pairs associated with the capital and located in relations from the Mikolov et al. (2013b) dataset, which contains a large collection of such relational pairs.

5.2.2 Other Wikidata Relations

We extracted concept pairs for each of the remaining 18 base relations from the Yuan et al. (2024) dataset Analogy KB, which provides curated concept pairs corresponding to a broad range of Wikidata relations.

5.2.3 Concept Pair Inclusion Criteria

For Sections 5.2.1 and 5.2.2, we established specific inclusion criteria. Since our ultimate goal is to utilize Wikipedia context with the two concept pairs in Section 5.3, we only included concept pairs that appeared together in at least five Wikipedia sentences and met a minimum threshold of relevance. For each base relation, we sampled up to 50 concept pairs; however, not all base relations had a sufficient number of pairs that satisfied the criteria. Consequently, the resulting dataset exhibits an imbalanced representation across different base relations. The total number of distinct concept pair was 584.

5.3 Retrieve all Relations for Concept Pairs

Having compiled 584 concept pairs, now our objective is to retrieve all other plausible relations between each pair, in alignment with our task formulation (see 4.1). For this, we follow a RAG (Yu et al., 2025) like approach. Given two concepts a_1 and a_2, first, we queried their respective Wikipedia pages (wik, 2025) and retrieved sequences mentioning both concepts. This retrieved sequence was then provided to GPT4.1 as context, prompting it to generate a list of potential relationships between a_1 and a_2. To ensure relevance, we manually filtered the results to remove duplicates (e.g. locatedIn, isIN), ambiguous relations (e.g. represents), opinion-based (e.g. isTheBest), transient relations (e.g. currentlyHappeningIn). The dataset contained a total of 573 distinct relations, with a mean of 5.53,

a median of 4, and a mode of 4. These statistics suggest that the distribution of relations per concept pair is approximately symmetric, with each concept pair associated with around five relations on average. Refer to Appendix 9.5 for the RAG prompts and 9.4 for the filtering process.

5.4 Ranking Relations

Now that we have all the relations $r_a \in R$ between two concept pairs a_1 and a_2, We ranked each relation as per the degree of it's importance to the concept pair. To achieve this, we employed a GPT4o supported Max-Diff Survey (Maximum Difference Scaling) (Louviere et al., 2015), a robust research technique for measuring preferences, priorities, and the relative significance of items. We divided the relation set r_a into n subsets, each of length $k : k < R$ such that each relation is equally present in the subsets. We selected $k = 4$ when $R \geq 4$ (total relations of a concept pair is more) and $k = 2$ when $R \in (4, 3)$. For $R \in (2, 1)$ we did not had multiple sets but a single query. These sets were given to GPT4.1 to label the most relevant and least relavant relations for each concept pairs. The Max-Diff survey therefore generates two probability distributions, one that depicts the probability that a particular relation r between a_1 and a_2 is the most relevant $P^m_{r \in r_a}(r)$. The other depicts the probability that a particular relation is likely to be least important $P^l_{r \in r_a}(r)$. This helps us score each relation for a concept pair and rank them accordingly. Therefore for a concept pair a_1 and a_2 the importance score of a particular relation becomes

$$s^a_i = P^m_{r_i \in r_a}(r_i) - P^l_{r_i \in r_a}(r_i) + \phi : \phi > 1$$

A positive offset ϕ is added to ensure non-negative scores, particularly for relations where the probability of the least important relation outweighs that of the most important one.

We used GPT-4.1 to rank relations using a Max-Diff survey framework because for its efficiency, reliability, and substantially lower cost than traditional crowd-sourcing. To validate this method, we compared LLM-generated rankings with those from human experts and crowd workers (Table 1). We evaluated Max-Diff responses for 12 concept pairs from the Country–Capital relation. Expert annotations were provided by seven domain experts familiar with the research. In parallel, we gathered responses from Amazon Mechanical Turk workers, filtered by a 95% HIT approval rate, at least 100 completed HITs, U.S. residency, English fluency, and passed attention checks. Results show that LLM rankings correlate more strongly with expert judgments than those from crowd workers, supporting the use of LLMs as a reliable and cost-effective solution for relational ranking in analogy tasks.

Method	Agreement with expert group
GPT-4.1	0.79
GPT-4o	0.71
GPT-4o-mini	0.62
GPT-3.5	0.53
M-Turk	0.56

Table 1: Agreement with expert group judgments in the MaxDiff survey: comparison between LLMs and MTurk workers.

5.5 Semantic Measures for Ground Truth

We incorporated three distinct measures to evaluate and compare the performance of LLMs.

5.5.1 Ranked Relation Overlap

This measure based on Christoph Lofi (2013)'s idea leverages the ranked relation set mined in Section 5.4. For each stem-target pair, we identify intersecting relations and use their ranked scores from the Max-Diff survey (see 5.4) as a proxy measure. The final score is obtained by summing the aggregate scores of all intersecting relations, with the highest-scoring pair selected as the best match. Formaly, from the notations in Section 4.3 and 4.4, let $r_{ab} = r_a \cap r_b$ be the overlapped relation between steam (a_1, a_2) and target (b_1, b_2). Let $S^{r_{ab}}_a$ and $S^{r_{ab}}_b$ be the scores corresponding to elements in r_{ab}, meaning:

$$S^{r_{ab}}_a = (S_a[i] | i \in r_{ab})$$

$$S^{r_{ab}}_b = (S_b[i] | i \in r_{ab})$$

Then compute the dot product of the restricted score vectors which represents the sum of element-wise products of scores for the elements that appear in both (a_1, a_2) and (b_1, b_2).

$$\sum_{i \in r_{ab}} S_a[i] \cdot S_b[i]$$

5.5.2 Context Embedding Similarity

Based on Turney et al. (2003)'s concept of *phrase vectors*, defined as a vector representation that captures the relational meaning between two concepts in a latent space. In our approach, we construct phrase vectors using sentences retrieved from

Wikipedia that contain the given concept pairs. As described in Section 5.3, we extract sentences from Wikipedia that mention both concepts for each stem and target concept pair. We then generate embeddings for these sentences using the openai's text-embedding-3-large model. The resulting sentence embeddings serve as the phrase vector representing the relational meaning between the concept pairs. Then we calculated the cosine similarity of the embeddings of the stem with each of the targets. Therefore, if $\{em_1^a, em_2^a, em_3^a..\}$ and $\{em_1^b, em_2^b, em_3^b..\}$ be the phrase vector of stem and a target. Then compute

$$\sum_{ij} Cosine(em_i^a, em_j^b)_{>0.5}$$

Since there are multiple sentences per pair, we retain only those with a similarity score greater than 0.5 and normalize the score by the number of sentences meeting this threshold. The correct answer is the target with the minimum distance from the stem. The threshold selection was guided by the cumulative distribution of similarity measures across all phrase vectors. For further discussed in Appendix 9.3.

5.5.3 Prototypical Similarity

This idea was first introduced as a sub-task of SemEval 2012 (Jurgens et al., 2012). Prototypicality refers to the extent to which a target concept is representative of the relation instantiated by the stem pair. Translating this notion into a computable measure is non-trivial; in this work, we approximate it by comparing the concatenated word embedding of the stem pair and its base relation with the word embedding of the candidate target pairs. For example, *London:England* representing capital is more prototypical of *Paris:France* than of *Ngerulmud:Palau*.

In practice, cities that occur more frequently across diverse corpora (e.g., news articles, encyclopedias) tend to be more prototypical. A city with high degree centrality—indicating stronger connectivity within relational networks can thus be considered more representative. Accordingly, raw word embeddings act as a proxy to encode such representativeness. For instance, *London* and *Paris* co-occur more often in text corpora than *London* and *Ngerulmud*. Since embedding models capture such co-occurrence patterns, we employ them as a proxy to approximate this semantic measure of prototypicality.

To quantify prototypicality, we have stem concept pair $a = (a_1, a_2)$ and its base relation r_{base} target concept pair $b_i = (b_{i1}, b_{i2})$. We compute the cosine similarity between

$$max_i(Cos(E(a, r_{base}), E(b_i)))$$

Where $E()$ is the embedding function. A higher similarity implies a stronger prototypical alignment between the stem its base relation and the target, while more orthogonal vectors suggest lower prototypicality.

5.5.4 Curation

Each of the 584 concept pairs was employed as a stem in a four-option multiple-choice analogy question. Target options were randomly selected from concept pairs within the same base relation group as the stem. For groups with a large number of concept pairs (e.g., *capital–country*), multiple instances were generated using the same stem, as the likelihood of duplicate targets was comparatively low due to the size of the group. In total, 1,210 test examples were constructed. For each instance, three ground-truth labels were defined, corresponding to three distinct semantic measures. Model performance was then evaluated against these three ground truths.

6 Evaluation

In this study, we have benchmarked several prominent generative Large Language Models (LLMs), including GPT-3.5, GPT-4 (OpenAI, 2023), LLaMA-2 (7B and 13B) (Touvron et al., 2023), LLaMA-3 (Grattafiori et al., 2024), Mistral-7B, Falcon-40B (Almazrouei et al., 2023), and Mixtral-13B (Jiang et al., 2023), etc. The evaluation of these models was conducted under three distinct prompting scenarios: zero-shot, automatic CoT Zhang et al. (2022), and knowledge-enhanced (Wijesiriwardene et al., 2024) prompting. In the context of zero-shot prompting, models were presented with analogy questions devoid of any prior context and asked to guess the correct answer. In the automatic chain-of-thought setting, analogy prompts were augmented by appending the directive *"think step by step"*, thereby eliciting explicit, stepwise reasoning from the model. In the knowledge-enhanced prompting setting, we explicitly guided models with reasoning directives tailored to each semantic measure. For the prototypicality measure, the prompt first introduced

the concept of prototypicality, emphasizing how certain targets are more representative exemplars of a relation than others, and instructed the model to select its response accordingly. For the ranked relational overlap measure, the prompt provided the set of salient relations associated with the stem pair. The model was asked to first infer the corresponding relations of each candidate target pair, then evaluate their relative rankings, identify the degree of relational overlap with the stem, and finally assign scores to the targets in order to select the most aligned candidate. For the context embedding similarity measure, the model was instructed to construct contextual text for both the stem and the candidate target pairs. It was then directed to compare these text and score the targets based on their semantic closeness to the stem. Please note that steering the model by prompts to compare embedding directly is non-trivial. This structured prompting framework ensured that the evaluation of each semantic measure was accompanied by a clear reasoning directive, allowing us to systematically probe whether LLMs can adapt their reasoning strategies in accordance with explicit semantic guidance. For each model prediction we also asked models to generate rationale to explain their answer choices.

7 Results and Discussion

7.1 Model Scale and Architecture

The results demonstrate a clear scaling advantage in analogy tasks. Smaller models such as Phi-2.7B, Gemma-2B, and RWKV-1.5B show consistently low accuracy across all semantic measures, rarely exceeding 20% in the Relational Overlap metric. In contrast, large-scale models such as LLaMA-3 70B, GPT-4o, and GPT-4.1 achieve accuracies well above 40% on embedding- and prototypicality-based measures. However, GPT-4o-mini surpasses 52% on prototypicality knowledge prompts, representing the highest overall performance in the benchmark. Which indicates model architecture might be a confounding factor for this anomaly. Architectural differences play a crucial role. RWKV models, despite competitive scale, consistently lag behind transformer-based families (LLaMA, Mistral, GPT). For example, RWKV-13B achieves only 22.7% relational overlap compared to 36% for LLaMA-3 8B. Similarly, Mistral-13B and LLaMA-2 13B exhibit comparable capacities, but both fall behind GPT-4 family models in embedding and prototypical measures. This underscores that architecture and pretraining corpus, not just parameter count, critically shape analogical reasoning performance.

7.2 Semantic Measures and Reasoning Strategies

Performance varied significantly across the three semantic measures:

- **Relational Overlap**: Even the strongest models underperformed, with GPT-4.1 and LLaMA-3 8B achieving only $\sim$30–36% accuracy. This suggests that capturing fine-grained structural relations remains a persistent weakness across architectures.
- **Context Embedding Similarity**: Models generally excelled on this measure, with LLaMA-3 70B (44.2%) and GPT-4o-mini (48.1%) leading. This indicates that LLMs rely heavily on distributional similarity in embedding space rather than systematic relation mapping.
- **Prototypicality**: Performance was highest overall, with GPT-4o-mini exceeding 52% and LLaMA-3 70B reaching nearly 49%. This reveals a strong tendency for LLMs to gravitate toward prototypical associations—choosing options that are frequent, salient, or canonical exemplars of a relation.

Taken together, the results show that while models are adept at leveraging surface-level semantic similarity and prototypical cues, they fail to robustly reason over deeper relational structures.

7.3 Impact of Knowledge Prompts

Knowledge-enhanced prompting improved results, but the gains were marginal and inconsistent. For Relational Overlap, improvements were modest (e.g., LLaMA-2 70B rose from 22.1% to 31.1%). By contrast, prototypicality saw larger boosts: Gemma-2B improved from 24.8% to 36.9%, and RWKV-1.5B rose from 30.0% to 35.2%. Context Embedding Similarity also benefited, with LLaMA-2 13B jumping from 28.3% to 33.8%. These findings suggest that explicit reasoning instructions help models align better with semantic criteria, though they cannot fully overcome structural reasoning deficiencies. It further invites evaluation of the dataset with more sophisticated knowledge-enhanced prompting strategies to examine whether such approaches yield performance improvements.

Models	$Rel - Overlap_{acc}$			$Context - Embedding_{acc}$			$Prototypicality_{acc}$		
	zero-shot	auto-cot	knowledge	zero-shot	auto-cot	knowledge	zero-shot	auto-cot	knowledge
Phi 2.7B	7.11	11.27	15.31	21.68	25.37	29.24	33.21	27.36	30.10
Gemma 2B	9.84	17.47	20.23	14.11	15.29	18.28	24.80	32.34	36.98
RWKV 1.5B	5.54	14.76	16.83	10.27	13.74	11.03	30.00	29.00	35.23
Gemma 7B	16.29	15.63	22.48	27.87	25.16	26.30	27.23	35.24	39.16
RWKV 7B	13.43	16.32	15.60	19.29	26.80	30.48	29.12	26.19	29.26
Mistral 7B	17.83	17.80	26.93	14.96	18.35	17.36	38.28	30.17	38.16
RWKV 13B	18.26	14.78	22.71	33.58	37.32	31.36	35.98	37.16	38.36
Mistral 13B	20.56	25.62	25.10	25.76	27.53	26.06	34.65	35.24	38.29
Llama2 13B	18.22	20.31	21.15	28.33	27.10	33.85	35.07	37.08	41.28
Llama2 70B	22.13	28.36	31.15	34.81	31.43	37.16	39.18	40.11	39.82
Llama3 8B	29.44	27.62	36.06	41.77	40.86	41.19	44.33	42.92	41.93
Llama3 70B	29.03	28.12	25.55	43.92	45.32	44.19	47.72	47.97	49.1 3
GPT 4o-mini	24.1 6	21.33	25.55	48.38	48.13	48.13	52.85	51.69	52.87
GPT 4o	27.95	28.68	30.02	41.77	42.10	43.83	46.48	45.65	44.25
GPT 4.1	29.61	29.19	30.93	39.78	40.69	38.46	41.60	42.50	39.86

Table 2: Accuracy of LLMs w.r.t semantic measures in our multiple-choice Analogy Task.

7.4 Summary of Insights

Overall, scaling and prompting enhance analogy task performance, but primarily by amplifying reliance on embedding similarity and prototypical associations rather than fostering genuine relational reasoning. GPT-4o-mini's strong results highlight that model efficiency combined with architectural design may outperform even larger models in certain contexts. However, the consistently low scores on relational overlap across all systems indicate a fundamental limitation of current LLMs: analogical reasoning remains shallow, biased toward salience and co-occurrence rather than structured mapping.

7.5 Rationale Analysis

In addition to accuracy, we examined the rationales generated by models when explaining their analogy choices. A recurring pattern was that model explanations often failed to reference the key relations that should have determined the correct answer. For example, when presented with a stem pair linked by a salient relation such as *capital of*, models frequently justified their choice by appealing to surface-level associations like geographic proximity or cultural similarity, rather than identifying all the other underlying structural relation as instructed by the knowledge prompts. This mismatch indicates that while models can produce fluent and plausible explanations, these rationales are not reliably grounded in the relational semantics of the task. Consequently, the generated rationales appear more as post-hoc justifications than evidence of genuine analogical reasoning, further underscoring the models' reliance on shallow heuristics instead of systematic relation-based inference.

8 Conclusion

This work introduced a novel benchmark designed to probe the analogical reasoning abilities of large language models in settings where multiple target solutions are semantically plausible. By moving beyond traditional single-answer analogy tasks, we demonstrated how LLMs respond when confronted with relational complexity and overlapping semantic structures. Our experiments revealed that while scaling and prompting strategies improve performance, LLMs overwhelmingly rely on prototypicality and distributional similarity rather than fine-grained relational reasoning. Even the strongest models consistently underperformed on relational overlap measures, underscoring a fundamental limitation in their ability to capture structured relational mappings.

These findings highlight an important gap between surface-level semantic fluency and genuine analogical reasoning. Although prompting techniques, such as knowledge-enhanced instructions, provide measurable gains, they fail to address the deeper structural deficiencies observed across models. This suggests that analogical reasoning in current LLMs is largely an emergent byproduct of scale and training data rather than the result of explicit relational abstraction.

Limitations

A limitation of this study lies in the complexity of mining all possible relations for any given concept pair, which is further compounded by the challenge of ranking these relations. Consequently, the study relies on an LLM-based Max-Diff survey. We addressed this by comparing the LLM Max-Diff survey with expert human and crowd-worker surveys

to justify our design choice. Another limitation of the study is the prompt sensitivity of decoder-based LLMs.

Beyond these, several additional limitations must be noted. First, although our dataset construction attempts to ensure semantic richness, the coverage of relations is still uneven across domains. Some relation categories, such as *capital–country*, are heavily represented, whereas others remain under-sampled. Second, the evaluation relies on three semantic measures-ranked relational overlap, context embedding similarity, and prototypicality which, while complementary, do not exhaustively capture the full spectrum of analogical reasoning strategies. Third, model performance differences are often marginal, making it difficult to draw strong conclusions about the superiority of one architecture or prompting strategy over another. Fourth, rationale analysis revealed that models frequently produce fluent but ungrounded explanations, suggesting that performance metrics alone may overestimate true reasoning ability. Finally, our study is limited to text-only LLMs; extending the framework to multimodal analogies involving images or diagrams could reveal different patterns of strengths and weaknesses.

Ethics Statement

The study focuses on testing existing publicly available LLMs without creating new models or enhancing potentially harmful capabilities. All experiments were designed to evaluate reasoning abilities in a controlled environment, avoiding deployment scenarios that could lead to unintended misuse or societal harm. To promote transparency and reproducibility, we have documented all methods, datasets, and evaluation metrics in detail. This study does not assume that LLMs truly "understand" analogies as humans do. Instead, their performance reflects patterns learned from data. We caution against over-interpreting results or deploying these models in sensitive areas requiring high reliability and accountability.

References

2025. Wikipedia, the free encyclopedia. https://en.wikipedia.org. Accessed: 2025-07-19.

Sonia Chernova Adrian Boteanu. 2015. Solving and explaining analogy questions using semantic networks.

Carl Allen and Timothy Hospedales. 2019. Analogies explained: Towards understanding word embeddings.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Christian Nieke Christoph Lofi. 2013. Modeling analogies for human-centered information systems.

Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. 2020. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, Online. Association for Computational Linguistics.

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

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,

Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang,

Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Douglas R. Hofstadter. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books, New York.

Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. In-context analogical reasoning with pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.

Filip Ilievski, Jay Pujara, and Kartik Shenoy. 2022. Does wikidata support analogical reasoning?

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

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

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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, Atlanta, Georgia. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Molly R. Petersen and Lonneke van der Plas. 2024. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance.

Shradha Sehgal, Bhavya, Krishna Phani Datta, Aditi Mallavarapu, and Cheng Xiang Zhai. 2024. Exploring ai-powered multimodal analogies for science education. *CEUR Workshop Proceedings*, 3840. Publisher Copyright: © 2024 Copyright for this paper by its authors.; 2024 Joint of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops, HEXED-L3MNGET 2024 ; Conference date: 14-07-2024.

Robyn Speer, Catherine Havasi, and Henry Lieberman. 2008. Analogyspace: reducing the dimensionality of common sense knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1*, AAAI'08, page 548–553. AAAI Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,

Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. 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, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. `https://www.wikidata.org`. Accessed: 2025-07-19.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

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.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Sreeram Vennam, Vinija Jain, Aman Chadha, Amitava Das, Ponnurangam Kumaraguru, and Amit Sheth. 2024. Knowledgeprompts: Exploring the abilities of large language models to solve proportional analogies via knowledge-enhanced prompting.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. 2022. AbductionRules: Training transformers to explain unexpected inputs. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 218–227, Dublin, Ireland. Association for Computational Linguistics.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

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.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

Hanzhang Zhou, Junlang Qian, Zijian Feng, Lu Hui, Zixiao Zhu, and Kezhi Mao. 2024. LLMs learn task heuristics from demonstrations: A heuristic-driven prompting strategy for document-level event argument extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11972–11990, Bangkok, Thailand. Association for Computational Linguistics.

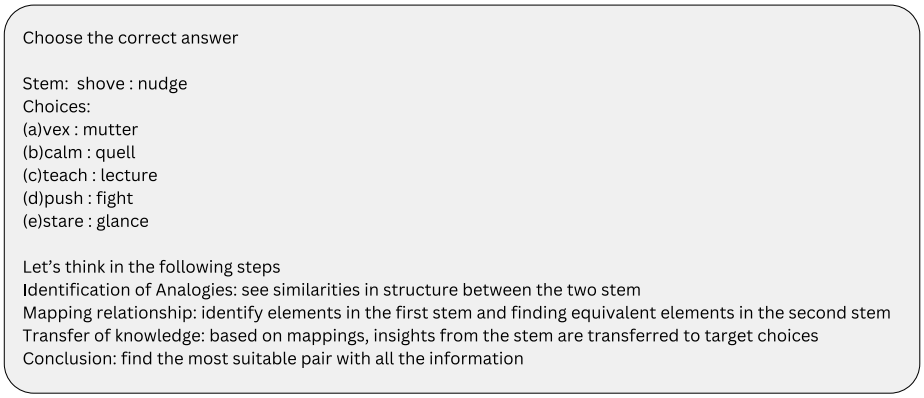
9 Appendix

9.1 Prompts for Section 3, SAT Benchmark

These are the prompts used to make the decoder model generate the correct answer to SAT(Turney et al., 2003) dataset.

9.1.1 Prompts for Oneshot Classification on SAT Analogy task with Chain of Thought Example

9.1.2 Prompts for Zeroshot Classification on SAT Analogy task



9.2 Example of base relations

- capital
- chairperson
- country of citizenship
- different from
- diplomatic relations
- filming location
- has contributing factor
- movement
- located in
- notable work
- operating system
- part of
- participant
- place of burial
- place of death
- separated from
- shares border with
- twin administered body
- work location
- worshipped by

9.3 Choice of Threshold in Wiki Embedding Measure Section 5.5

This section provides a detailed analysis of the rationale behind the chosen threshold. Specifically, we analyze the distribution of similarity scores across all phrasal vectors derived from our dataset.

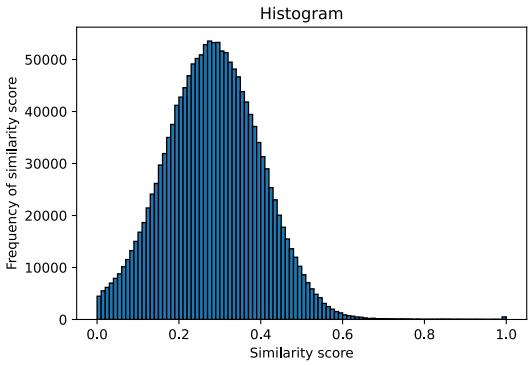


Figure 6: Histogram of similarity scores reveals a skewed distribution towards low scores

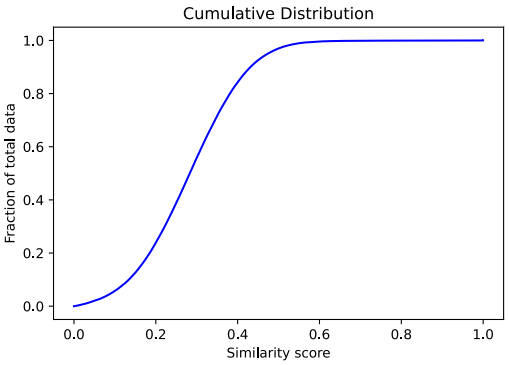


Figure 7: CDF reveals that around 0.4 to 0.6 we have high end of the distribution

The results presented in Figure 6, Figure 7, and Table 3 indicate that the distribution of similarity scores is highly left-skewed. Consequently, a large proportion of the data exhibits low similarity around 0.3, while only a small fraction exceeds a similarity score of 0.6. This supports the choice of a 0.5 threshold, which retains a balance of both highly similar and moderately similar sentences.

Similarity Score (Sim)	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Percentage of Data $\geq Sim$	100	94.23	76.10	44.17	15.71	**3.01**	0.40	0.11	0.06	0.04

Table 3: The proportion of data samples exceeding a given discrete similarity score. E.g., for similarity score ≥ 0.3 we will have 44.17% of the data

9.4 Filtering Relevant Relations

After extracting relevant relationships from Wikipedia and LLMs, we conducted a manual filtering process to refine the quality of the relations, categorizing them retracted relations into four distinct groups.

- **Ambiguous Relation:** An ambiguous relation refers to a connection between two entities whose nature is open to multiple interpretations, often leading to potential disagreement regarding its precise meaning. E.g., London ***represents_the*** United Kingdom, Money ***is*** Power, Tokyo ***leads*** Japan.

- **Opinion Based Relation:** An opinion-based relation is a relationship between two entities that is based on subjective judgment, personal preference, or cultural perception rather than objective facts. E.g., Paris ***is_ the_most_romantic_city_in*** France, Tokyo ***is_ the_most_exciting_city_in*** Japan

- **Transient Relation:** A transient relation is a temporary, non-permanent association between two entities, where the connection is subject to change due to external factors such as time. We filterd only those transient relations which have already changed and no longer valid, E.g., Pluto ***was_classified_as*** planet, The Sun ***was_thought_to orbit*** earth.

- **Duplicate Relation:** The relations that could be merged to keep a single relation and discard the others. E.g., when we have Tokyo ***is_in*** Japan and Tokyo ***is_located_in*** Japan, we can discard ***is_in***.

9.5 Prompts for Relation Generation

The prompt used for Relation generation is as follows
system: "You are an assistant that can extract relationships between two words from a paragraph provided as context"

user: extract the relationships between {**concept**} and {**concept**} from the context below {**wikipedia-context**}return the relations in {**style**} format

style can be predicate, rdf, knowledge graph. This paper used rdf as the style

9.6 Prompts for zero-shot generations for Solving the New Analogy task

This figure displays the prompts used to generate GPT model predictions for the analogy tasks.

system prompt: You are an assistant that can solve 4-way multiple-choice SAT analogy task
user prompt: Which one among the following targets best matches the stem ? only one option is correct.

Stem: New York : United States

Targets:
(a)Baghdad : Iraq
(b)Kabul : Afganisthan
(c)Shanghai : Chaina
(d)Beijing : China

return the target pair only and no other text

Figure 8: This is the prompt used to generate GPT guess for the correct target. system prompt and user prompts are not a part of the prompt but a field in openai chat completion

9.7 Knowledge Enhanced Prompts for Solving our Analogy task

These reasoning hints were used to generate GPT model predictions for the analogy tasks pertaining to Knowledge enhanced performance.

9.7.1 Prototypicality

Hint: Identify the relationship between the given stem pairs. Then, choose the option that best represents this same relationship. Select the option that is the most prototypical example of the relation. Context

9.7.2 Context Embedding

Hint: First, generate context between the stem pairs and all the options Score each option 0–1 for how well its context matches the stem context. Break ties by preferring the option whose directionality $(X ß Y)$ matches the stem.

9.7.3 Relational Overlap

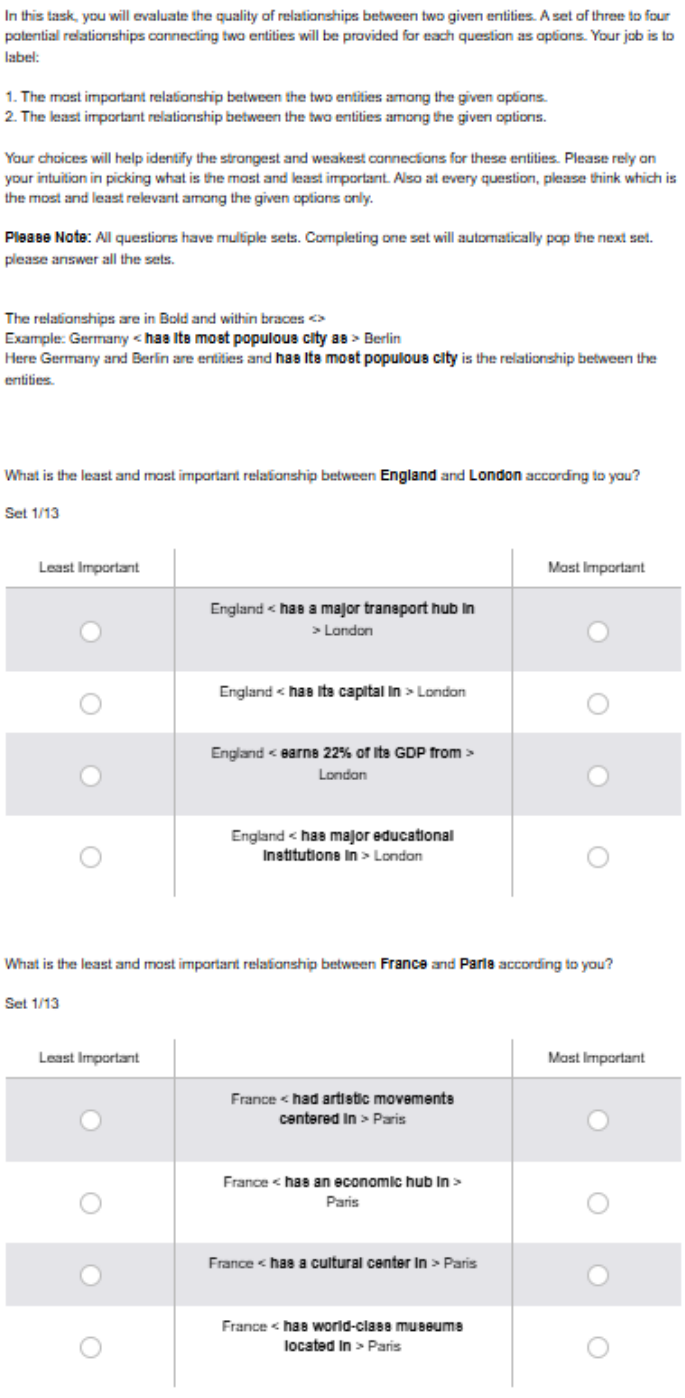
Hint: The relations between the stem pairs are given relations the numbers (0-1) represents the importance of the relations for the stem Find yourself all relations of the option pair. Score the relations as per their degree of importance to the parent option Find for each option how well its relations overlaps with the stem context. Score each option by multiplying the overlapped relation score with the stems relation score and adding all the overlapps together Write this down in latex format,

9.8 Max-Diff Survey

A dummy link to a sample Max-Diff survey used in this paper is provided in this link. A dummy response can be completed https://www.surveyking.com/survey/ons79. Also snapshots are provided in this section

9.8.1 Max-Diff Survey Template

To assess the relative importance of different relational features, we conducted a Max-Diff (Maximum Difference Scaling) survey (see Section 5.4). Participants were presented with sets of relational pairs and asked to select the most and least representative relation in each set.



In this task, you will evaluate the quality of relationships between two given entities. A set of three to four potential relationships connecting two entities will be provided for each question as options. Your job is to label:

1. The most important relationship between the two entities among the given options.
2. The least important relationship between the two entities among the given options.

Your choices will help identify the strongest and weakest connections for these entities. Please rely on your intuition in picking what is the most and least important. Also at every question, please think which is the most and least relevant among the given options only.

Please Note: All questions have multiple sets. Completing one set will automatically pop the next set. please answer all the sets.

The relationships are in Bold and within braces <>
Example: Germany < **has its most populous city as** > Berlin
Here Germany and Berlin are entities and **has its most populous city** is the relationship between the entities.

What is the least and most important relationship between **England** and **London** according to you?

Set 1/13

Least Important		Most Important
○	England < **has a major transport hub in** > London	○
○	England < **has its capital in** > London	○
○	England < **earns 22% of its GDP from** > London	○
○	England < **has major educational institutions in** > London	○

What is the least and most important relationship between **France** and **Paris** according to you?

Set 1/13

Least Important		Most Important
○	France < **had artistic movements centered in** > Paris	○
○	France < **has an economic hub in** > Paris	○
○	France < **has a cultural center in** > Paris	○
○	France < **has world-class museums located in** > Paris	○

9.8.2 Max-Diff Survey Sample Results

This figure presents the aggregated responses from a single batch of Max-Diff survey, where participants rated the most and least representative relational pairs in each set. The collected responses allow for a quantitative assessment of the perceived importance of different relations with probability scores and votes per relation. The distribution of responses provides insight into consensus patterns and variability in human judgment, informing our evaluation of relational similarity measures. In the paper responses from LLM was used and not human. This is just for demonstration

Answer	Share of Preference	Probability*	Distribution	Least Important	Most Important
France < has its capital in > Paris	57.79%	94.61%		0	25
France < has its most populous city as > Paris	7.55%	69.63%		1	12
France < has its most visited city as > Paris	6.41%	66.06%		1	10
France < has its largest city as > Paris	5.48%	62.49%		3	10
France < has world-class museums located in > Paris	5.48%	62.49%		3	10
France < has its fashion capital in > Paris	4.39%	57.14%		3	7
France < has major landmarks located in > Paris	3.54%	51.78%		6	7
France < has a cultural center in > Paris	2.85%	46.43%		7	5
France < had artistic movements centered in > Paris	1.83%	35.72%		9	1
France < has an economic hub in > Paris	1.56%	32.16%		12	2
France < hosts the Tour de France that finishes in > Paris	1.32%	28.59%		13	1
France < earns 32% of its GDP from > Paris	1%	23.23%		16	1
France < has a city > Paris	0.81%	19.66%		17	0