ANALOBENCH: Benchmarking the Identification of Abstract and Long-context Analogies

Anonymous ACL submission

Abstract

Humans regularly engage in analogical thinking, relating personal experiences to current situations (X is analogous to Y because of Z). Analogical thinking allows humans to solve problems in creative ways, grasp difficult concepts, and articulate ideas more effectively. Can language models (LMs) do the same? To answer this question, we propose ANALOBENCH, a benchmark to determine analogical reasoning ability in LMs. Our benchmarking approach focuses on aspects of this ability that are common among humans: (i) recalling related experiences from a large amount of information, and (ii) applying analogical reasoning to complex and lengthy scenarios. We test a broad collection of proprietary models (e.g., GPT family, Claude-v2) and open source models such as LLaMA2. As in prior results, scaling up LMs results in some performance boosts. Surprisingly, scale offers minimal gains when, (i) analogies involve lengthy scenarios, or (ii) recalling relevant scenarios from a large pool of information, a process analogous to finding a needle in a haystack. We hope these observations encourage further research in this field.[1]

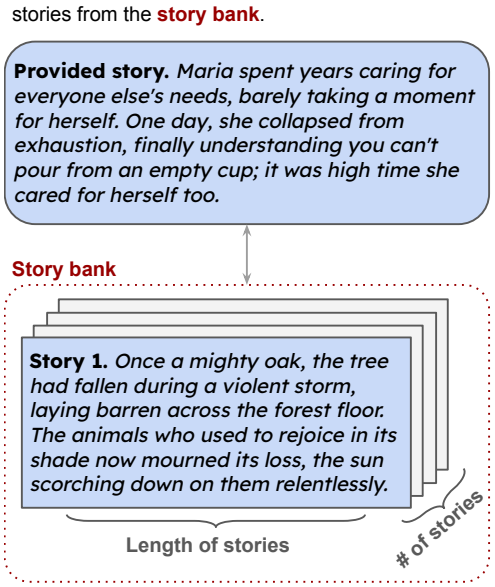


Figure 1: The problem setup: given a story, the goal is to identify an analogous story from a story bank. We study the difficulty of this goal for LMs by varying the following parameters: (i) length of stories, (ii) number of stories in the story bank. In the example, both "Maria" and "the oak" lose the ability to provide for others. While the strength of analogies can vary, we design our benchmark to account for this variation.

1 Introduction

Analogy is the ability to think about relational patterns (Holyoak et al., 2001) and forms an integral aspect of human communication (Hofstadter, 2001; Gentner and Hoyos, 2017). This cognitive ability helps humans understand new or difficult concepts by relating them to more familiar experiences (Holyoak and Thagard, 1996). Analogical thinking plays a critical role in some of the major breakthroughs in human history, such as the discovery of gravity or even Einstein's theory of relativity (Hesse, 1965; Stepan, 1986; Hofstadter and Sander, 2013). It was this very analogy-driven progress that Newton aptly described as *"standing upon the shoulders of giants,"* itself an analogy. If modern language models (LMs) (OpenAI, 2023; Touvron et al., 2023) can leverage analogical thinking, then we can expect wide-ranging implications for future tasks.

We assess the ability of modern LMs to handle components of analogy making. Two important features characterize how humans form analogies in creative pursuits. (1) Humans are able to pinpoint analogies between prolonged experiences (e.g. "obtaining a PhD is like running a marathon"). (2) Humans can recollect relevant analogs from a large collection of past experiences to form analogies (Keane, 1987; Wharton et al., 1994; Han et al., 2018). To what extent are LMs capable of such abilities?

To answer the above questions, we introduce

[1]We have linked our dataset for the reviewer's attention should they like to verify it manually.

ANALOBENCH, a benchmark for analogical reasoning over natural language stories that convey abstract concepts with varying level of difficulty. While the dominant treatment of analogies has been limited to word-level lexical analogies[2] (Mikolov et al., 2010; Pennington et al., 2014), we instead focus on analogies defined on natural language documents, such as the one shown in Figure 1. In the example, the central figure of each stories (Maria / the "mighty oak") loses the ability to provide for the needs of others ("collapsed from exhaustion" / "the tree had fallen"). The use of stories as components of analogies provides a natural way to introduce abstract relational patterns. In total, we collect 340 pairs of high-quality analogous stories from human annotators after multiple rounds of review and editing.

As Figure 1 shows, we are broadly interested in quantifying the extent to which LMs are capable of identifying analogous stories from a given pool of candidate stories, similar to humans' ability to recollect past experiences and relate them to new situations.

We characterize this goal with two tasks (§3.3). First, we consider a setup where the pool is limited to a few stories. Among these few candidates, the model is expected to select exactly one story as the closest analogy to a given story ($\boldsymbol{T}_1$). Good performance requires demonstrated ability in identifying complex analogies, assuming a small pool of candidates. In our second task, we maintain a large (≈ 200) pool of candidate stories ($\boldsymbol{T}_2$) — in performing well on this task, a model will have demonstrated ability in identifying analogies from long-context memory.

Additionally, we explore how well performance scales with length. We are inspired by the remarkable ability of humans to abstract over long and elaborate stories, and leverage such abstractions to identify analogies. By evaluating our proposed tasks on longer stories, we measure the extent LMs can abstract over complexities of longer stories. In practice, we repeat each experiment with the same stories told using $\approx$ 1 sentence, 10 sentences, and 30 sentences.

We benchmark existing open-source and private language models to measure their ability to identify abstract and long-context analogies (§4). We find that, while scaling LMs leads to better performance in 1-sentence stories, the gains afforded by scale is minimal for longer stories. Furthermore, the gap between humans and GPT4 is 6.9% on 1-sentence stories, but increases to 28.8% on 30-sentence stories, demonstrating that long and complex analogies pose a challenge for LMs.

In summary, we provide: (1) ANALOBENCH, a novel benchmark with two tasks to determine the analogical reasoning capabilities of LMs, (2) a recipe for identifying analogies using LMs that, to the best of our knowledge, is the first to consider the role of story length and the recollection of relevant stories, and (3) a thorough analysis of analogical reasoning ability in a wide range of state of the art language models.

2 Related Work

We review the body of relevant work, including datasets and modeling results.

Analogical reasoning datasets. Various efforts have attempted to build analogical reasoning benchmarks. Within the AI literature, the majority of these works focus on lexical analogies (*i.e.*, *"man" to "woman" ≈ "boy" to "girl"*) (Sternberg and Nigro, 1980; Turney, 2008; Green et al., 2012; Jurgens et al., 2012; Mikolov et al., 2013b,c; Gladkova et al., 2016; Lu et al., 2019; Ushio et al., 2021). Most of these datasets are created manually, although there are also lexical analogy resources that are created semi-automatically. For example, Yuan et al. (2023) presents a dataset with over a million lexical analogies derived from a knowledge base of subject-object-verb triplets. Our work ventures beyond lexical analogies and focuses on analogies that involve paragraphs of raw-form text, without any assumptions on their structure.

Another group of datasets are from cognitive science, some of which involve long sentences. These datasets were originally intended to be used for the study of analogical reasoning in humans (Gick and Holyoak, 1980; Wharton et al., 1994; Keane, 1987; Gentner et al., 1993; Weinberger et al., 2016). The majority of these datasets are too small to provide reliable benchmarking for models. Among these GENTNER (Gentner and Toupin, 1986) contains 54 instances and was created to examine the development of systematicity (i.e., sensitivity to parallels based on more complex relations). Recently, (Webb et al., 2023) observes strong performance of LLMs on these datasets, which motivates introducing more challenge analogical reasoning benchmark.

[2] e.g. "rock" is to "solid" as "water" is to "liquid"

A concurrent work is Jiayang et al. (2023) who introduce STORYANALOGY, a benchmark of 24K sentence pairs, which were generated semi-automatically using GPT-3 and then relabeled by human annotators. Compared to this work, our benchmark is much smaller as we prioritize data quality over size. Our seed data is all written by humans, at the cost of size, mainly because we aimed at effective evaluation. Furthermore, we evaluate the effectiveness of models to solve analogies in long paragraphs.

It is worth noting that there is also a literature on *visual* analogies (Sadeghi et al., 2015; Bitton et al., 2023; Reed et al., 2015; Zhang et al., 2019) that is different from the scope of this work. Interested readers can refer to Ichien et al. (2020) who provide a thorough review of the prior datasets both in computer science and cognitive science literature.

Analogical reasoning in LMs. Since the rise of pre-trained LMs, we have witnessed remarkable gains in the abilities of these models in tackling analogical reasoning (Ichien et al., 2023; Webb et al., 2023). Bhavya et al. (2022) studied the ability of LMs (the GPT3 family) in generating analogous statements with prompting by literal mentions of "analogy" in prompts. Through crowdsourcing experiments, they observe that the then largest models (e.g., `davinci`) were able to generate analogies that matched the quality of human-generated analogies. Another remarkable milestone is reported by Webb et al. (2023) who evaluate GPT3 on various analogical reasoning tasks (Raven's standard progressive matrices, letter string analogies, etc.) and report that "GPT-3 displayed a surprisingly strong capacity for abstract pattern induction, matching or even surpassing human capabilities in most settings." Our results also confirm such progress—the abilities of LMs indeed increase with the improvements in the scale. However, our benchmark reveals major limitations of LMs that was not easily observable in the prior work (e.g., the weakness of LMs in solving analogies that involve longer inputs).

3 ANALOBENCH: A Benchmark for Abstract and Long-Context Analogies

We describe desired attributes of benchmarking analogies (§3.1), discuss the construction of ANALOBENCH (§3.2) and the tasks devised based on this dataset (§3.3).

3.1 Design Considerations

We discuss on the unique qualities of analogical reasoning to guide and motivate our design:

Assess the breadth of analogies. The universe of analogies is vast, and any LM is likely only able to predict a small (often easy) subset of this universe. While measuring the precision of LMs is important, an ideal benchmark should also measure their recall (how well they capture deep and abstract analogies). Generative evaluation might not fully capture this depth, as there may exist many analogies that the LM cannot predict. To assess what an LM cannot predict, we propose a set of analogies of our own choosing, and evaluate analogical identification on this set (§3.3).

Assess the relative strength of analogies. The quality of analogies inherently lie on a spectrum—some are stronger and some are weaker. Ideally, a task formulation should capture the continuum of analogy strengths. We thus frame our analogical identification task more specifically as a ranking task, where the best answer must be preferred over incorrect choices.

Use creative analogies. LMs perform worse on creative (i.e. rare) data. Thus a benchmark that aims to challenge LMs should feature analogies that are creative. To that end, we introduce novel and diverse, human-written analogies created through a semi-randomized process (§3.2).

3.2 Dataset Creation

① Curating analogical sentence-pairs. We collected 340 analogies from 4 human annotators (the authors) after multiple rounds of editing. The human annotators included native English speakers and non-native speakers who all attended university in the United States. These analogies then served as true positives in our experiments. These annotations of analogies are arranged in pairs of sentences that share similar relational patterns. For example, in Figure 2 the sentences "He danced off his sugar high then promptly fell asleep" and "The weather finally became pleasant following the stormy week" form an analogy. While these two sentences are topically dissimilar (dancing vs weather), they nevertheless share abstract relational patterns.

The construction of these sentence pairs follows this process: For each annotation, a random sentence is provided to the annotator, who is tasked with creating a corresponding analogous sentence.

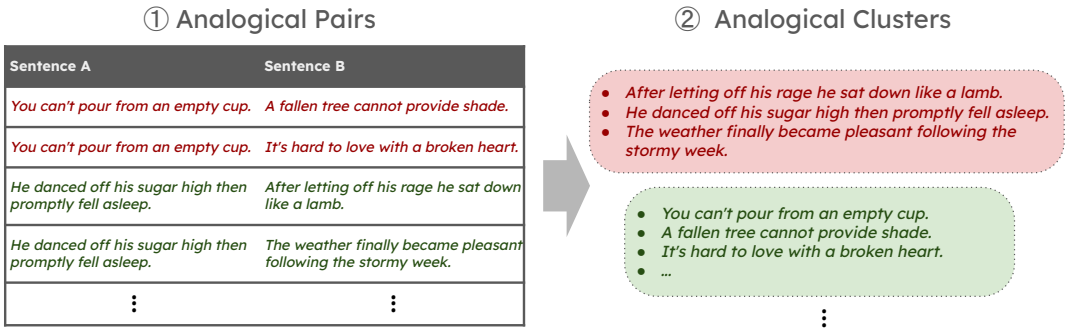


Figure 2: An overview of dataset creation (§3.2). ① **Left:** Human annotators are asked to create pairs of analogous sentences. Sentences can be repeated from analogy to analogy. ② **Right:** Pairs that share a sentence can be grouped into a cluster of mutually analogous sentences by transitivity.

Source sentences were sampled from Cambridge Dictionary examples of idioms found on an online resource[3] and a dataset of metaphors (Bizzoni and Lappin, 2018) filtered down to keep only examples with the strongest or second-strongest grades. To encourage innovative and abstract analogies, the annotator is given 3 random words to incorporate in the newly formed sentence.[4] During our pilot study, the introduction of random words was found to induce more creative annotations.

There are guidelines that the annotators adhere to. Firstly, they are instructed to avoid using similar sentence structure, topics or words as the original sentence. This is to eliminate any easy shortcuts that might allow LMs to recognize an analogy without having identified relational patterns. For example, an LM might mistakenly use similar phrasing between a pair of sentences to detect an analogy. Instead, the analogy between two sentences should be established on the basis of shared relational patterns. Finally, a separate reviewer scrutinizes the contributed sentence pairs to ensure their clarity, accuracy, and effective use of analogy.

② **Forming analogical clusters.** Our collected data is structured such that the same sentence can appear in several analogous sentence pairs. This allows us to organize our dataset into sets of analogous *clusters*, where all pairs of sentences in a cluster are mutually analogous by transitivity. Each cluster is manually inspected and adjusted to confirm mutual analogousness. Furthermore, different clusters that happen to share common relational patterns are combined. We then use these clusters to setup the tasks in §3.3.

③ **Analogy elaboration.** To investigate the effect of story length on the complexity of analogies, we collect elaborated versions of each story. First, longer stories requires analogical reasoning over longer contexts, a task which scales in difficulty for LMs, as shown by the recent results (Chen et al., 2023; Khandelwal et al., 2018; Liu et al., 2023). Second, the longer the stories in an analogy, the more room for expressing abstract relational patterns.

To implement this elaboration, we use GPT-4 to convert sentence-level analogies into detailed stories with a target length of 10 sentences and 30 sentences. We selected GPT-4 for its advanced story generation capabilities and proficiency over other LMs in generating coherent and complex text.[5] To balance creativity and logical consistency, we configured the model with parameters such as temperature $= 1$ and top-$p = 0.95$. We provide the prompt templates used in Appendix D. Although we later evaluate GPT-4 on its own generations, we demonstrate that self-evaluation bias does not affect our conclusions by testing GPT-4 on stories generated by a different model (§5.1).

Statistics. Table 1 shows the overall statistics of our collected analogical clusters and their elaborations. Overall, we compiled a total of 340 stories grouped into 47 clusters. On average, each cluster consists of about 7.2 stories.

3.3 Analogy Identification Tasks

With the clusters of analogies defined (§3.2), we leverage this data to devise two tasks to benchmark the capability of state of the art LMs at analogical reasoning. In §1, we introduced two components of analogy making. Each task aims to evaluate both components in conjunction. Given a query story, both tasks involve identifying analogous stories to the given one from a story bank. In the first task, we

[3] See this link.

[4] Randomization was achieved by sampling nouns, verbs, and adjectives from here.

[5] In our pilot experiments, we compared the elaborations using GPT-4, PaLM and Claude, and ultimately chose GPT-4 because of its accurate yet creative elaborations.

Measure	Value
# of clusters	47.0
avg. size of clusters (stories)	7.2
avg. length (sentences) of 1 sentence stories	1.2
avg. length (sentences) of 10 sentence stories	11.9
avg. length (sentences) of 30 sentence stories	31.2
avg. length (subwords) of 1 sentence stories	21.3
avg. length (subwords) of 10 sentence stories	225.8
avg. length (subwords) of 30 sentence stories	552.8

Table 1: Summary of dataset statistics. The dataset consists of 47 clusters with an average of 7.2 stories each, and stories vary in average sentence and subword length.

maintain a small story bank to focus the challenge on rating a few candidates, thereby disentangling it from the challenge of long-context reasoning. In the second task, given a story, a model must identify analogous stories from a large story bank. We intend this approach to be analogous to how humans recollect and employ their past experience to form analogies.

T_1: Identify analogies from mini story bank. This task confronts the model with choosing the most fitting analogy from 4 options. Given a sentence or story, the model must select the most suitable option from a lineup consisting of one correct answer and three intentionally crafted distractors to assess its discernment of analogical relationships. Each answer choice is prefixed by a letter from [A, B, C, D] (e.g. "D. A fallen tree cannot provide shade"). We prompt each model to answer the question: "Which of the following is the most analogous story to the target story?" To guide the LM to make a selection, we impose an additional condition in the prompt that the generation must be one of the four letters. More details of our approach can be found in Appendix E.

An example of this task is shown in Figure 5. Note, given the elaborated stories discussed in (step ③ in §3.2), we have three datasets of multiple-choice questions for each story length (1-sentence, 10-sentences, 30-sentences).

T_2: Identify analogies from large story bank. In this task, given a story, the model must identify the top 10 most analogous stories from a carefully assembled, fixed story-bank consisting of 200 different stories. Each story is prefixed by a number from 1 to 200 (e.g. "1. Kim checked the papers..."). For this task, we prompt each model to generate a list of integers representing the index numbers of its selections. Following this, we employ precision and recall metrics to analyze its performance. More details and examples are provided in Appendix G.[6]

Like the earlier task, we study this task in three distinct setups as a function of story length (1 sentence, 10 sentences, 30 sentences). The size of the story-bank provided to the model varies considerably on different datasets. For 1-sentence dataset, the story-bank for it contains 4K tokens. For 30-sentence story-bank, it contains 110K tokens. The long-context nature of this task poses a major challenge for LMs. Due to these constraints, our evaluation of T_2 (§4.3) is limited to the few models (GPT-4 and Claude-v2) that can handle long-context.

4 Main Experiments

We structure our experimental assessment around two primary tasks aimed at evaluating the analogical reasoning of LMs. We discuss the experimental setting including the metrics, models and human evaluation (§4.1), followed by the results.

4.1 Experimental Setting

Metrics. All scores are reported as percentages. For T_1 (analogies from a *small* story bank, §3.3) we use accuracy as the primary measure of success. Each example has multiple candidate analogies. A solver gets a score of 1 if it chooses the most analogous story and $1/k$ if it reports no-answer or a k-way tie that includes the correct answer ($k = 4$ in our dataset.) For T_2 (analogy from a *large* story bank, §3.3), we report common retrieval metrics such as Mean Average Precision (MAP), Precision@K, Recall@K, and Mean Reciprocal Rank (MRR) (Manning et al., 2008).

Evaluated models. To determine the analogical reasoning abilities of the LMs, we include an assortment of high-profile models in our benchmarks. The lineup features GPT-4 (OpenAI, 2023), GPT-3.5 (Brown et al., 2020), LLaMA2-chat (Touvron et al., 2023), UnifiedQA (Khashabi et al., 2020, 2022), XwinLM (Xwin-LM, 2023), WizardLM (Xu et al., 2023), Tulu2 (Ivison et al., 2023), Zephyr (Tunstall et al., 2023), Claude-v2 (Anthropic, 2023). To minimize variations in model

[6]We also considered using the LM to assign likelihoods to analogous stories, then ranking the entire story-bank by likelihood. However, the extent to which modern LMs are well-calibrated remains unclear, especially in this domain. We conducted preliminary studies that attempted to score the strength of an analogy between two sentences. Scores were wildly inconsistent between runs and different in-context examples, even on low temperature settings. The factors that contribute to the inconsistent behavior remain unclear, and thus we do not define our task in this manner.

responses, we set the decoding parameters to temperature $= 0.3$ and top-$p = 0.95$.

Of the models considered, we do not include supervised baselines, where a model is trained or finetuned on labelled analogy data. We believe that learning analogical thinking from supervised training is misguided. A training set of analogies that can comprehensively approximate the entire universe of analogies is unlikely to exist. Therefore, while a model can fit *some* set of analogies, that model might fail to generalize to out-of-distribution examples or even other analogy datasets.

Human evaluation. We conducted human evaluation to measure whether the task is well-defined and has a reasonable quality. This process was meticulously applied to our T_1 task (Analogy Selection, §3.3) across different levels of complexity: 1-sentence, 10-sentence, and 30-sentence scenarios. To make the 30-sentence task more manageable, the annotations were done for 30 instances. However, for 1-sentence and 10-sentence settings, we annotated 50 instances.

For each level of complexity, we enlisted three additional annotators[7] (who were not involved in the dataset construction) to evaluate the analogy scenarios. Each annotator began by selecting their personal answer choice without conferring. This exercise led to high-agreement among the annotators (results in Appendix C).

Following this individual judgment phase, disagreements were adjudicated via discussion among the annotators. During these discussions, the annotators were encouraged to exchange their rationales behind their initial selection and converge upon one collective answer that we used for evaluation.

We did not run human annotations for T_2 due to the immense reading load expected of annotators. Additionally, since the two tasks are based on the same set of labeled data, we focus our human annotations on T_1 to establish the quality of the presented data.

4.2 Result: Mini Story Bank (T_1)

We benchmark how well our models can identify analogies from a mini story bank (so as to disentangle this task from other challenges associated with long-context reasoning). Our results are reported in Table 2 and Figure 3. More detailed results are reported in Table 5 of the Appendix. Overall, our analogical reasoning benchmark challenges state of the art language models.

LMs do not outperform humans. Our benchmarking reveals that analogical ability varies widely among modern LMs. While many models perform non-trivially (i.e. better than 25% accuracy achieved by random guessing), and some models such as GPT-4 perform considerably well, no model is able to outperform humans in any setting. Among open-source models, the largest models (70B) dominate the results for the shortest story length setting, with the exception of UnifiedQA which is supervised with different data than the rest of the models.

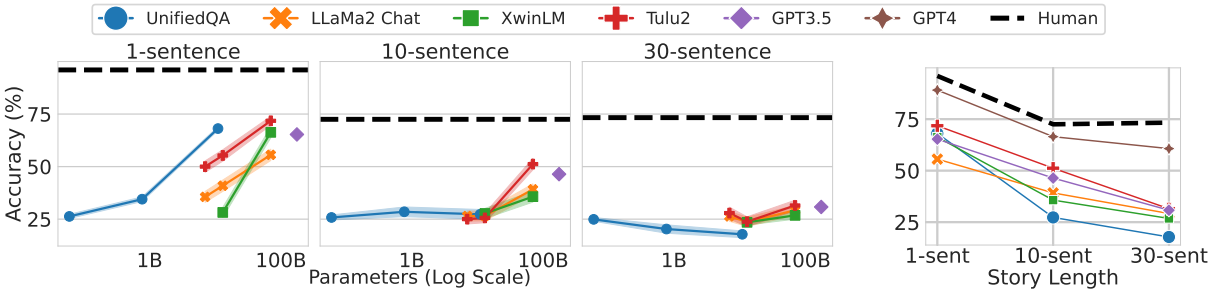
	Model ↓ - Story length →	1-sent	10-sent	30-sent
	Random	25	25	25
Open-source	Zephyr (7B)	55.1	27.1	20.3
	UnifiedQA (11B)	68.1	27.3	17.8
	WizardLM (13B)	41.1	29.1	25.7
	LLaMA2-chat (70B)	55.6	39.2	29.5
	XwinLM (70B)	66.3	35.7	26.8
	Tulu2 (70B)	**71.8**	**51.2**	**31.5**
Private	Claude	68.2	30.2	25.9
	GPT3.5	65.3	46.4	30.8
	GPT4	**89.1**	**66.5**	**60.7**
	Human	***96.0***	***72.5***	***73.3***

Table 2: Benchmarking various models for T_1 (§4.2). For open-source models, we only show the results of the largest available sizes in their model family. **While the best models perform somewhat close to human in short analogies (1-sentence), the human-AI gap increases in longer stories**.

Analogy length degrades LM performance. We evaluate our lineup of models on stories consisting of 1, 10, and 30 sentences. In Figure 3b all models exhibit degradation as story length increases. In contrast, while human performance also decreases with longer length, their performance stops decreasing for 10- and 30-sentence stories. Thus, the performance gap between humans and LMs increases with longer context-length. These results to suggest that analogical reasoning over longer inputs poses an inherent challenge for LMs.

Model scaling benefits are limited on long stories. Even if performance diminishes with increased story length across all models, as long as performance improves with model size, a sufficiently large model can solve this problem. To test this possibility, we evaluate models of varying sizes within the UnifiedQA, LLaMA2, XwinLM, and

[7]These annotators (also authors) share the same demographic as the other annotators in §3.2 and were not aware of the experimental design during annotation



(a) Results with varying model scale. The error margins are based on the standard error. While scaling LMs is effective among short (1-sent) stories (left), **the benefit of scale is negligible for longer stories** (middle and right).

(b) With increasing story length, **model accuracy decreases, while their gap with humans increases**.

Figure 3: Accuracy of LMs on T_1 (§4.2).

Tulu2 families on T_1. Our results in Figure 3a indicate that while performance scales with LLM size in the single sentence setting, we do not observe the same trend in longer settings. Specifically, in longer stories performance plateaus across model family as model size increases. The observed trend indicates a limit to the benefits of scaling model size when handling complex analogices.

4.3 Results: Large Story Bank (T_2)

Having evaluated our lineup of models on the mini story-bank setting, we now turn our attention to the full story-bank setting. As stated in §3.3, fitting the full story bank in the context window requires us to consider only long-context models such as GPT4-Turbo and Claude. In this experiment, given a story, each model must identify the top k most analogous stories from the story-bank. We report the precision-recall curves for $k = 1, ..., 10$ in Figure 4 and provide further details in Table 6 of the Appendix.

LM performance approaches random. We evaluate both models as well as a trivial baseline where k random documents are retrieved. Both models perform similarly to the trivial baseline in most cases. The only exception is the performance of GPT4-Turbo in the single-sentence setting. While impressive, the performance of GPT4-Turbo is nevertheless near trivial in lengthier settings. These evaluations test the limits of the best modern LMs. If humans can recollect relevant experiences to form analogies (Keane, 1987; Wharton et al., 1994), then our results suggest that further research is necessary to achieve parity in LMs.

5 Further Analysis

5.1 Evaluating Self-Generated Stories

In past experiments, we utilized GPT-4 to extend single-sentence stories into versions containing 10 or 30 sentences. Consequently, the relatively high accuracy of GPT-4 may stem from evaluating its own generated content. To address this, we compare the performance of GPT-4 on its own stories compared to stories generated by Claude. As a baseline, we measure the performance of Claude on its own stories and stories generated by GPT-4. We report our results in Table 3 and find that GPT-4 encounters negligible performance degradation upon switching to Claude generations. Additionally, GPT-4 consistently outperforms Claude on Claude generations. These results suggest that the relatively high performance of GPT-4 is likely attributed to factors other than evaluating its own generations.

Eval. \ Gen.	10 Sentences		30 Sentences	
	Claude	GPT4	Claude	GPT4
Claude	36.5	30.2	33.5	25.9
GPT4	69.7	66.5	57.6	60.7

Table 3: Perf. of different evaluators and generators on 10- and 30-sentence stories (§5.1). **GPT-4 performance experiences minimal change when evaluating Claude generations.**

5.2 Effect of Dataset Error

We cannot rule out the possibility that our dataset may contain errors. Indeed, whether incorrect predictions from humans are attributable to human error or dataset error is unclear. To reduce the likelihood of dataset error affecting our conclusions, we deem analogies that were correctly predicted by humans to be relatively free of error, and repeat experiment T_1 on those analogies. We test GPT-

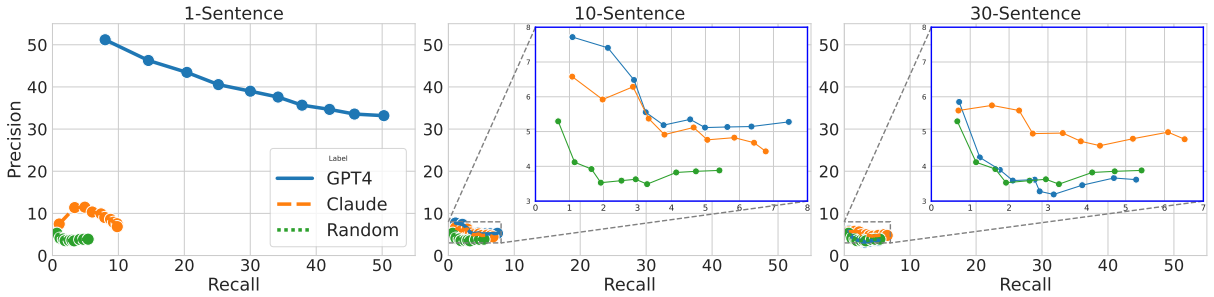


Figure 4: Precision-recall plot (in percentage) of LMs on T_2 (§4.3) at three different story lengths (1, 10, 30 sentences). **With increasing story length, the precision-recall of the models approaches random.**

4, Tulu2, and XwinLM and report our results in Table 4. All trends reported in §4.3 are still observed in this low-error setting, suggesting that our conclusions are unlikely to be affected by dataset error.

Model	T_1: Classification (accuracy%)		
	1 sentence	10 sentences	30 sentences
GPT4	91.7	58.3	22.7
Tulu7	47.2	30.1	16.7
Tulu13	52.8	30.1	22.2
Tulu70	74.0	34.7	25.0
Xwin13	27.8	19.4	22.9
Xwin70	42.2	31.9	29.1

Table 4: Model accuracy on true positive human predictions in T_1 (§5.2) at three different story lengths (1, 10, 30 sentences). **All trends are consistent with the original task.**

5.3 Longer Analogies are Easier for Humans

In §3.2 we hypothesized that analogy length corresponds to complexity. While our results clearly indicate that longer analogies pose a greater challenge for LMs, perhaps a more interesting question is whether they pose a greater challenge for humans. Surprisingly, human annotators found the 30 sentence setting easier than the 10 sentence setting, observing that the added details in the longer setting aid in disambiguation when performing the task. This opinion is supported by annotator performance in Table 2.

6 Discussion

Limits of modern LMs in analogical thinking. A clear consensus on whether LMs can adequately perform analogical thinking has remained elusive. While some find that LMs are proficient analogical reasoners (Webb et al., 2023; Ichien et al., 2023), others have challenged this notion (Jiayang et al., 2023). Throughout our experiments, we repeatedly find that modern LMs display limited ability to engage in key aspects of analogical thinking. Crucially, performance does not scale with model size on longer stories. Unlike the LMs evaluated, humans can identify analogies between even the longest stories with relative ease. These observations clearly suggest that LMs lack some key mechanism to think analogically. Overall, our results establish the need for further research to encourage analogical thinking in LMs.

Downstream applications and future work. What downstream applications can we expect from analogically reasoning LMs? We discuss examples to illustrate the potential of analogical LMs. In science, analogy provides a source of inspiration for innovation. For instance, the design of artificial neural networks was inspired by biological neural networks. An analogy driven scientific search engine would accelerate such innovation, allowing researchers to consider relevant ideas across vastly different contexts (Hope et al., 2017). In law, Zou et al. (2024) has represented consistency in legal decisions as an analogical reasoning problem, where decisions in a current case should follow that of an analogous case. An analogical search engine would aid in the identification of relevant cases. Given these wide-ranging applications, we hope that our findings motivate future work towards equipping LMs with better analogical reasoning capabilities.

7 Conclusion

Analogical reasoning is an important aspect of human cognition, with wide-ranging potential for future research. To benchmark this ability in LMs, we define a general approach by scaling the length of stories and the context from which they need to be retrieved. Our benchmark exposes the limitations of analogical reasoning in modern LMs. We release ANALOBENCH to motivate further research.

8 Limitations

In our experiments we benchmark many models. While trying more models and performing additional prompt-engineering could have affected results, in the end we were constrained by the available computing resources. Additionally, we cannot exclude the possibility that LMs encountered labelled analogies during training or finetuning, especially proprietary models such as GPT-4. While our dataset is more challenging than existing ones, it comes with various simplifying assumptions and cannot capture the potentially-infinite range of analogies. Future work should extend the existing datasets to capture more complex forms of analogical reasoning.

9 Ethics Statement

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. The work presented here does not immediately raise any ethical concerns, to our knowledge. Beyond the scope of this work, analogical reasoning should be applied with care, otherwise due to its inherent subjectivity it may potentially lead to misleading or incorrect conclusions.

References

Safa Alsaidi, Amandine Decker, Puthineath Lay, Esteban Marquer, Pierre-Alexandre Murena, and Miguel Couceiro. 2021. A neural approach for detecting morphological analogies. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Anthropic. 2023. Model card and evaluations for claude models.

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy generation by prompting large language models: A case study of instructgpt. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 298–312.

Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. 2023. Vasr: Visual analogies of situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 241–249.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* (NeurIPS).

Jaime G Carbonell. 1983. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*, pages 137–161. Elsevier.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Catherine A Clement and Dedre Gentner. 1991. Systematicity as a selection constraint in analogical mapping. *Cognitive science*, 15(1):89–132.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2018. Towards understanding linear word analogies. In *Annual Meeting of the Association for Computational Linguistics* (ACL).

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

Dedre Gentner and Christian Hoyos. 2017. Analogy and abstraction. *Topics in cognitive science*, 9(3):672–693.

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

Dedre Gentner and Cecile Toupin. 1986. Systematicity and surface similarity in the development of analogy. *Cognitive science*, 10(3):277–300.

Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology*, 12(3):306–355.

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.

Adam E Green, David JM Kraemer, Jonathan A Fugelsang, Jeremy R Gray, and Kevin N Dunbar. 2012. Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(2):264.

Ji Han, Feng Shi, Liuqing Chen, and Peter RN Childs. 2018. A computational tool for creative idea generation based on analogical reasoning and ontology. *AI EDAM*, 32(4):462–477.

Mary Hesse. 1965. Models and analogies in science. *British Journal for the Philosophy of Science*, 16(62).

Douglas R Hofstadter. 1984. The copycat project: An experiment in nondeterminism and creative analogies. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

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

Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books.

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

Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *ACM Conference Knowledge Discovery and Data Mining* (KDD), pages 235–243.

Nicholas Ichien, Hongjing Lu, and Keith J Holyoak. 2020. Verbal analogy problem sets: An inventory of testing materials. *Behavior research methods*, 52:1803–1816.

Nicholas Ichien, Dušan Stamenković, and Keith J Holyoak. 2023. Large language model displays emergent ability to interpret novel literary metaphors. *arXiv preprint arXiv:2308.01497*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint 2311.10702*.

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

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.

Mark Keane. 1987. On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology*, 39(1):29–41.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQA-v2: Stronger Generalization via Broader Cross-Format Training. *arXiv preprint arXiv:2202.12359*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP) - *Findings*.

Matthew Lamm, Arun Chaganty, Christopher D Manning, Dan Jurafsky, and Percy Liang. 2018. Textual analogy parsing: What's shared and what's compared among analogous facts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 82–92.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Hongjing Lu, Ying Nian Wu, and Keith J Holyoak. 2019. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181.

C. D. Manning, P. Raghavan, H. Schütze, et al. 2008. *Introduction to information retrieval*. Cambridge university press Cambridge.

Esteban Marquer and Miguel Couceiro. 2023. Solving morphological analogies: from retrieval to generation. *arXiv preprint arXiv:2303.18062*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations* (ICLR).

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Annual Conference of the International Speech Communication Association* (INTERSPEECH), pages 1045–1048.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. 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.

Pierre-Alexandre Murena, Marie Al-Ghossein, Jean-Louis Dessalles, Antoine Cornuéjols, et al. 2020. Solving analogies on words based on minimal complexity transformation. In *IJCAI*, pages 1848–1854.

OpenAI. 2023. Gpt-4 technical report.

Robert Oppenheimer. 1956. Analogy in science. *American Psychologist*, 11(3):127.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP), pages 1532–1543.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. 2015. Deep visual analogy-making. *Advances in neural information processing systems*, 28.

Fereshteh Sadeghi, C Lawrence Zitnick, and Ali Farhadi. 2015. Visalogy: Answering visual analogy questions. *Advances in Neural Information Processing Systems*, 28.

Roger C Schank. 1999. *Dynamic memory revisited*. Cambridge University Press.

Nancy Leys Stepan. 1986. Race and gender: The role of analogy in science. *Isis*, 77(2):261–277.

Robert J Sternberg and Georgia Nigro. 1980. Developmental patterns in the solution of verbal analogies. *Child Development*, pages 27–38.

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. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint 2310.16944*.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912.

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 *Annual Meeting of the Association for Computational Linguistics* (ACL).

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

Adam B Weinberger, Hari Iyer, and Adam E Green. 2016. Conscious augmentation of creative state enhances "real" creativity in open-ended analogical reasoning. *PloS one*, 11(3):e0150773.

Charles M Wharton, Keith J Holyoak, Paul E Downing, Trent E Lange, Thomas D Wickens, and Eric R Melz. 1994. Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, 26(1):64–101.

Patrick H Winston. 1980. Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Team Xwin-LM. 2023. Xwin-lm.

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*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327.

Xinrui Zou, Ming Zhang, Nathaniel Weir, Benjamin Van Durme, and Nils Holzenberger. 2024. Reframing tax law entailment as analogical reasoning.

Supplemental Material

A Additional Related Work

Here we cover additional related work that did not fit in the main text.

Analogical reasoning before LMs. The research on analogical reasoning in AI and cognitive science for the longest time has focused on four-term analogies (Hesse, 1965) (e.g., "Baltimore to Maryland is like NYC to New York"). In the era of symbolic AI era, an extensive literature focused on engineering symbolic systems that processed analogical reasoning (Winston, 1980; Carbonell, 1983; Hofstadter, 1984; Schank, 1999). These works focus on richer representation for alignment of analogous symbols and their dynamic retrieval from a memory structure.

The more complex the analogies are, the more complex representation they require (Holyoak et al., 2001). Naturally, it meant that solving the analogy problem require solving the representation problem. The increasing progress in extracting representations of language led to more progress in analogical reseasoning. A decade ago, the earlier generation of representation learning algorithms such as Word2Vec (Mikolov et al., 2010, 2013a) famously showed linguistic regularities equivalent to lexical analogies (Pennington et al., 2014; Ethayarajh et al., 2018) Thereafter, a large body of works focused on effective ways of eliciting analogies from word embeddings (Murena et al., 2020), sometimes through neural networks or symbolic reasoning frameworks built atop these embeddings (Lamm et al., 2018; Alsaidi et al., 2021; Marquer and Couceiro, 2023).

Analogical reasoning in humans. The cognitive ability to process analogies likely has been with homosapiens since the time they developed their languages, as evidenced by written Babylonian or Egyptian relics (Holyoak and Thagard, 1996). These written documents convey a variety of ideas: friendship and emotions, dangers and enemies, power and greed, and so on.

Analogies also made their way to science. Greeks used analogies to describe their understanding of physical concepts, such as sound waves spreading like water waves. Physicists used similar abstractions to understand light waves by formulating analogies to known physical waves, leading to "wave theory of light". Analogies are so prevalent in scientific development that renowned physicist J. Robert Oppenheimer called it an "indispensable and inevitable tool for scientific progress" (Oppenheimer, 1956).

Cognitive science is the community which adopted a scientific and systematic treatment of analogical reasoning in human cognition. Within cognitive science, analogical reasoning was viewed as mental models that utilize structure alignment via relations (Gentner, 1983; Clement and Gentner, 1991). Analogical reasoning was also studied under pragmatic contexts such as the goal of the environment or the problem solving (Gick and Holyoak, 1980). Hofstadter (2001); Gentner and Hoyos (2017) argue that analogical reasoning is the "core of cognition".

B Visualizing the Task

The following Figure 5 shows an overview of ANALOBENCH, for both the story expansion ③ and the task creation §3.3. Our abstract analogy identification benchmark features two tasks: (T_1) Identifying analogies from mini story bank and (T_2) Identifying analogies from mini story bank. Each task is repeated at varying story lengths ($\sim$ 1, 10, and 30 sentences), with GPT-4 extending each story to target length. We find that while analogical reasoning shows signs of emergence, reasoning over longer and more complex analogies remains a challenge for state of the art LMs.

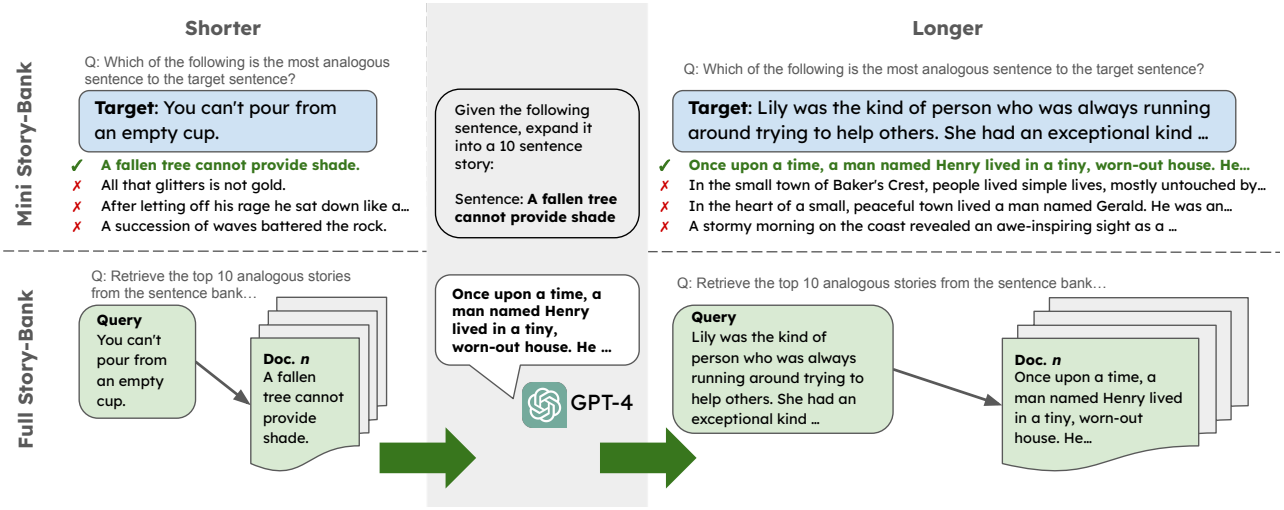


Figure 5

C Inter-annotator Agreement Among Human Annotators

Using our task definition, when we measure the inter-annotator agreement on our human-written analogies (1-sentence), we find that all three human evaluators agree unanimously on 47 of 50 analogies. The high degree of inter-annotator agreement is a quantitative indicator of our dataset's objective evaluation and quality.

Task	Agreement Rate
1-sentence	0.94
10-sentence	0.70
30-sentence	0.73

D Further Details on Analogy Elaboration

We expand a single sentence to craft a story spanning 10 or 30 sentences. This directive applies to both GPT-4 in §3.2. Below is an example:

Example prompts provided to GPT-4 for story elaboration:

Input:

```
Given the following sentence, expand it into a 10-sentence story.
Sentence: All that glitters is not gold.
```

Output:

```
In the small town of Baker's Crest, people lived simple lives, mostly untouched by the allure of
wealth and luxury. The quiet town was buzzing one day with the news of a travelling merchant who
claimed to sell "real gold" artifacts. The townspeople gathered at the market square, attracted
by the glinting objects displayed on his cart.
The items glittered under the sunlight, their sheen unmatched, making everyone believe they were
indeed made of gold. The merchant, a persuasive man, told captivating tales of the artifacts
being lost treasures of ancient kings. An air of greed and desire blew over the town, as everyone
wanted a piece of this assumed wealth. They traded their hard-earned savings for the glittering
items, believing they had struck a fortune.
It was not until a wise old man from the town examined the artifacts closely and declared them
to be mere brass, cleverly polished to resemble gold. All the glitter had indeed fooled them,
reinforcing the adage, "All that glitters is not gold." The realization dawned, leaving the
townspeople embarrassed, having learned their lesson the hard way.
```

Throughout this creative process, we regulate it with a temperature setting of 1 and a top_p value of 0.95. We experimented with different temperatures, but these adjustments introduced additional issues. A high temperature caused the narrative to diverge from the core meaning of the original sentence, whereas a low temperature led to repetitive elements which rendered generated stories highly similar due to shared analogous traits.

Assessing the quality of story expansion. We conduct an experiment to test the ability of GPT-4 to extend stories while hewing to the original source. If GPT-4 is successful, then the original source (hypothesis) must entail from the extended story (premise). Modern LLMs are understood to be highly performant on the textual entailment task. Thus, we use the recently-released Claude-3 to predict entailment, taking care to avoid any potential bias in these evaluations that might unfairly favor the generations of GPT-4. As baselines, we randomly pair the premise and hypothesis for the 10- and 30-sentence setting.

Story Comparison	Entailment Rate
1 vs 10 sentence (random)	0.01
1 vs 10 sentence	0.95
1 vs 30 sentence (random)	0.03
1 vs 30 sentence	0.97

We show that nearly all our source stories entail from the extended versions.

E Prompts Used for Evaluating LMs for T_1

Figure 6 demonstrates the adaptation of a basic prompt to run various model evaluations for T_1 task. We begin with the basic prompt and adjust it slightly to comply with the specific instructions of each model, as depicted in the second tier of the diagram. The third tier presents examples of responses generated by the models. Also, we set the temperature=0.3 and top_p=0.95 for all of the model evaluations.

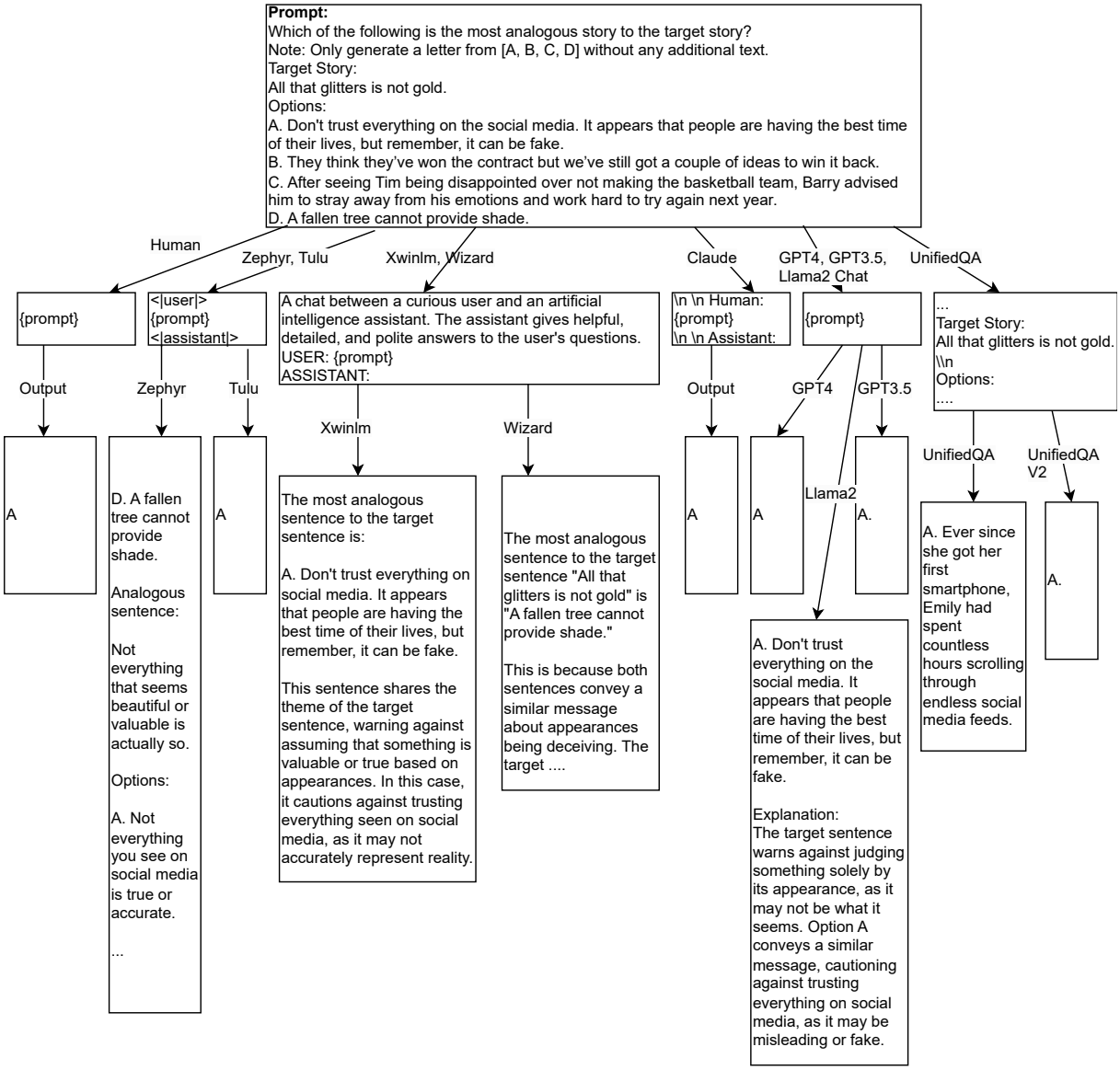


Figure 6: Analogy Selection Prompt for Different Models

F Detailed Results for T_1

Table 5 in our research paper presents the comprehensive set of results from our T_1 experiments discussed in §4.2. We assessed the abilities of numerous open-source models as well as GPT-4 and Claude-v2 on this particular task. We use 4xA100 to evaluate all of the models.

Model	Number of params	T_1 Analogy Selection (accuracy)		
		1 sentence	10 sentences	30 sentences
Random	–	25%	25%	25%
UnifiedQA	11B	68.1%	27.3%	17.8%
UnifiedQA v2	11B	53.8%	29.1%	23.6%
LLaMA2-chat	7B	35.6%	26.5%	26.3%
LLaMA2-chat	13B	40.9%	26.5%	23.7%
LLaMA2-chat	70B	55.6%	39.2%	29.5%
XwinLM	13B	28.2%	27.7%	23.5%
XwinLM	70B	66.3%	35.7%	26.8%
WizardLM	13B	41.1%	29.1%	25.7%
Tulu2	7B	50.0%	25.0%	27.9%
Tulu2	13B	55.3%	25.6%	23.8%
Tulu2	70B	71.8%	51.2%	31.5%
Zephyr	7B	55.1%	27.1%	20.3%
GPT3.5	175B	65.3%	46.4%	30.8%
GPT4	?	**89.1%**	**66.5%**	**60.7%**
Cluade	?	68.2%	30.2%	25.9%
Human	–	96.0%	72.5%	73.3%

Table 5: Performance of different models on analogy selection tasks.

G Prompts used for evaluating LMs for T_2

In §3.3 we discuss T_2, which is identifying the top 10 most analogous stories from a fixed bank of 200 stories. The following example shows the detail of the prompt.

GPT-4 Model Input and Output

Input:

Retrieve the top 10 analogous stories from the sentence bank for the following target story:
NOTE: Only generate an index number without any additional text. For example: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Target Story:
All that glitters is not gold.
Sentence Bank:
1.Kim checked the papers in a rush so that she can have more free time. But, now she needs to redo them as half of the class complained.
2. Liam lied to get into the school; Lary did not. Liam had a difficult time trying to hide the deception as a result. But unlike Liam, Lary did not have to worry about anything else, so he had a terrific time.
3. I am sorry, but you would now have to present your work before you can go for the vacation.
4. A fallen tree cannot provide shade.
5. He is the winner of three Grammy awards for god's sake! People consider him to be the god of rap.
...
197. It is not that cold today, but I'd still go by car since I can't afford to get sick.
198. I do not want to spoil your mood but I have to babysit my nephew today.
199. Every cigarette you smoked is a threat to your health.
200. He knocked the nail into the wall with a hammer.

Output:

4, 14, 29, 59, 97, 111, 113, 137, 172, 188

H Detailed Results for T_2

Table 6 in our research paper presents the comprehensive set of results from T_2. We assessed the abilities of GPT-4 Turbo and Claude-v2 on this particular task(§4.3). We use 4xA100 to evaluate all of the models. Here are some detailed results of it:

Metrics	T_2: [GPT4-turbo]			Claude-v2			Random	Oracle
	1 sentence	10 sentences	30 sentences	1 sentence	10 sentences	30 sentences		
P@3	42.9%	6.5%	3.9%	10.2%	5.3%	5.4%	3.9%	100%
P@5	38.5%	5.2%	3.6%	8.9%	4.6%	4.7%	3.6%	100%
R@3	20.1%	2.9%	1.8%	4.3%	2.4%	2.2%	1.6%	48.9%
R@5	29.7%	3.8%	2.7%	6.6%	3.6%	3.2%	2.5%	81.6%
MAP	55.4%	14.2%	10.8%	6.3%	1.9%	3.4%	1.7%	100%
MRR	64.2%	15.6%	11.3%	18.9%	9.8%	13.4%	11.1%	100%

Table 6: Performance metrics for T_2 using and Claude-v2 at different sentence lengths.

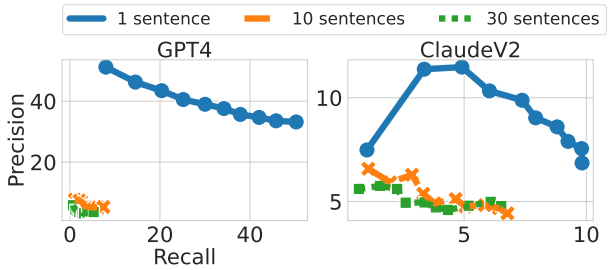


Figure 7: The figures indicate that GPT-4 and Claude-v2 excel in the task of retrieving 1 sentence, but their performance decreases with the retrieval tasks of 10 sentences and 30 sentences.

Calculation of 'Random' and 'Oracle' Baselines In the context of the table above, precision and recall calculations involve two lists of integers: "result" and "golden." In typical precision and recall computations, the "result" list is derived from the models' generations. However, for random calculations, the "result" list consists of integers from 1 to 10. This choice is influenced by our prompt: "NOTE: Only generate an index number without any additional text. For example: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10". Specifically, for challenging tasks, GPT-4 and Claude tend to generate a list ranging from 1 to 10 based on this prompt as default. The random calculation is then performed using this list. In the case of the Oracle calculation, we designate the "result" list to be the same as the "golden" list.

I Experiment: Evaluation on Different Stories Lengths for a Fixed Total Context Window

In our earlier experiments in §4.2 and §4.3 upon changing the length of each story, we also change the length of the total prompt (i.e., the concatenation of all the stories in the story bank). This essentially creates a confounding two variables that impact the difficulty of the tasks for LMs: (i) length of each story; (ii) the total length of the context. To address this confounding variable, here we fix (ii) and vary (i).

We fix a total context window length budget. Specifically, we fix this budget to be 2K and 1.5K tokens. Then, we fit as many stories that would fit within this total context window budget. The number of the stories that fit in the context window are shown in Table 7.

Total Context Length	Number of stories			**Scaled** Accuracy			Accuracy		
	1-sent	10-sent	30-sent	1-sent	10-sent	30-sent	1-sent	10-sent	30-sent
1500	72	6	3	0.04	0.15	0.07	0.05	0.29	0.38
2000	100	10	4	0.01	0.03	0.08	0.02	0.13	0.31

Table 7: Merged performance metrics for predictions across varying context lengths and story lengths for Tulu2 70B, with and without scaled accuracy.

We report the accuracy values for these evaluations, but these values are not comparable to across different length since they have different lower-bounds. For example, a story bank of size 3 leads to a lowerbound of 1/3, while the lowerbound for a story bank with 72 stories is 1/72.

Besides the accuracy metric, we also report a scaled accuracy. The scaling is necessary here to make sure that the numbers are all ranged from 0 to 100. To scale a given accuracy value x, we can plug it in the following formula:

$$\text{scaled-acc} = \frac{x - \text{random-acc}}{1 - \text{random-acc}},$$

where random-acc $= 1/(\text{size of story bank})$. Overall the results of scaled accuracy values in Table 7 are small. Essentially all of our stories of varying length remain difficult, even after accounting for a fixed context window size. Whether story length is a stronger factor or the context window length remains somewhat inconclusive and requires more future work.

J Using Claude-v2 for Story Elaboration

Similar to §3.2, in §5.1 we expand a single sentence to craft a story spanning 10 or 30 sentences with Claude-v2 this time. This directive is similar to how we prompt GPT-4 (example shown in §D), albeit with a slight modification in the guidance given to Claude-v2. Below is the instruction:

Input:

```
\n\n Human:
Given the following sentence, expand it into a 30-sentence story.
Sentence: All that glitters is not gold.
\n\n Assistant:
```

Output:

```
Here is a 10-sentence story expanding on the sentence "All that glitters is not gold":
Lisa was browsing the jewelry store, looking for a new necklace. So many of the pieces caught
her eye with their glittering gems and shiny metals. She tried on a beautiful gold necklace set
with bright green emeralds that seemed to glow. Although it was very pretty, when she looked at
the price tag she gasped - it cost a small fortune! Lisa realized that just because something
glitters, doesn't mean it's as valuable as it appears. She put the necklace back and continued
browsing for something more affordable. In a corner case, she found a simple silver chain with
a small pearl pendant.  It didn't sparkle as brightly, but she knew pearls were classic and
elegant. Lisa tried on the necklace and loved how it looked on her. She decided it was perfect,
and happily purchased it for a fraction of the price of the emerald necklace. Though it didn't
glitter quite as much, she knew this beautiful pearl pendant was the true treasure.
```