

Zero-shot Comparison of Large Language Models (LLMs) Reasoning Abilities on Long-text Analogies

Kara Combs
Air Force Research
Laboratory
Kara.combs.1@us.af.mil

Trevor J. Bihl
Air Force Research
Laboratory
Trevor.bihl.2@us.af.mil

Spencer Howlett
Applied Research
Solutions
showlett@appliedres.com

Yuki Adams
Applied Research
Solutions
yadams@appliedres.com

Abstract

In recent years, large language models (LLMs) have made substantial strides in mimicking human language and coherently presenting information. However, researchers continue to debate the accuracy and robustness of LLMs' reasoning abilities. The reasoning abilities of thirteen LLMs were tested on two long-text analogy datasets, named Rattermann and Wharton, which required them to rank a series of stories from most analogous to least analogous compared to a source story. On the Rattermann dataset, GPT-4 obtained the highest accuracy of 70%. As a whole, LLMs seem to struggle with over-emphasizing similar story entities (characters and settings) and a lack of awareness of higher-order relationship(s) between stories. LLMs struggled more with the Wharton dataset, with the highest accuracy achieved being 46.4% by GPT-4o, and all but nine LLMs performing below random chance accuracy. Although LLMs are improving, they still struggle with higher-cognitive tasks such as analogical reasoning.

Keywords: large language models, analogical reasoning, zero-shot learning

1. Introduction

Through the release of ChatGPT in November 2022, the democratization of large language models (LLMs) occurred and a wide variety of advanced AI capabilities entered the public marketplace and public consciousness. However, considerable work before ChatGPT occurred in foundational research involving statistical language models (LMs) and natural language processing (NLP), such as n-gram approaches (Niesler & Woodland, 1996); with the advent of deep neural networks, neural-based LMs appeared in the 2010s, such as Word2vec (Goldberg & Levy, 2014), but they were limited to solving basic NLP problems (Zhao, et al., 2023). With the development of transformer-based algorithms, the ability to solve problems with context-awareness came about (Zhao, et al., 2023). From this

background, in 2020, GPT-3 (Brown, et al., 2020), the precursor to the models behind ChatGPT, was released by OpenAI.

From this foundation, and due to demonstrated LLMs' abilities to function as useful virtual assistants, an explosion of models began appearing beginning in 2022 (Raiaan, et al., 2024; Minaee, et al., 2024; Combs, Bihl, & Ganapathy, 2024). Following in the footsteps of OpenAI, many tech companies and organizations have launched their LLMs. Since then, some of the most prominent LLMs include OpenAI's GPT-4 & GPT-4o, Meta's Llama family of models (Meta AI, 2023; Meta, 2023; Meta, 2024), and Google's PaLM, Gemini, and Gemma models (Narang & Chowdhery, 2022; Google, 2023; Gemma Team, Google DeepMind, 2024). The research and commercial landscapes are now rich with fine-tuned versions of open-source LLMs and comparisons on a variety of tasks such as question-answering, text generation, and summarization (Hadi, et al., 2023; Liusie, Manakul, & Gales, 2024).

LLMs have shown impressive performance on many evaluation metrics (see (Chang, et al., 2024)) and standardized tests (see (Bommineni, et al., 2023; Liu, et al., 2023; Katz, Bommarito, Gao, & Pablo, 2024)), which corresponds to the bottom of Bloom's Taxonomy shown in Figure 1, associated with lower cognitive processes. Thus, in the public consciousness, this has resulted in many believing that LLMs are intrinsically intelligent (Dodgson, 2023; Mitchell & Krakauer, 2023). However, LLMs' reasoning abilities are highly debated (Webb, Holyoak, & Lu, 2023; Hodel & West, 2023). In humans, reasoning requires the foundational lower cognitive processes, but places a much greater emphasis on the middle-to-higher cognitive processes in Bloom's Taxonomy (Kurtz, Gentner, & Gunn, 1999).

This debate and limited understanding of LLM's higher cognitive abilities, specifically reasoning, inspired the following research questions:

RQ1. How do current state-of-the-art LLMs compare based on their higher cognitive abilities?

RQ2. How do we evaluate LLMs' higher cognitive performance on multiple-choice reasoning questions?

RQ1 is answered through our analysis of LLM performance on two long-text analogy problems, which tests their analogical reasoning ability. Holyoak, & Lu (2023) previously looked at the analogical reasoning abilities of GPT-3 and 4 on a variety of analogical reasoning tasks. However, we have broadened the model scope to an additional twelve LLM and narrowed the scope to long-text analogy problems. To answer RQ2, we propose two metrics. First, we look at the selection accuracy of the most analogous option. Second, we consider the average rank accuracy of the most analogous option.

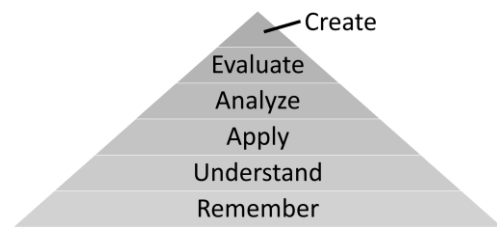


Figure 1. Bloom's Taxonomy (modified from (Anderson & Krathwohl, 2001))

First, we begin with background information on general LLM reasoning and analogical reasoning abilities within LLMs in Section 2. In Section 3 we describe our methodology related to LLM & hyperparameter(s) selection, selection of the long-text analogy dataset, and the LLM prompt templates, respectively. Metrics of interests are described in Section 4 and in Section 5, we present our comparison results and discussion thereof. Finally, Section 6 is our conclusion.

2. Background

LLMs are machine learning algorithms, often Transformer-based, that have a multitude of parameters (often hundreds of billions or more) and are trained on large text databases (Zhao, et al., 2023). Key to LLMs are transformer-based methods, including the encoder (e.g., BERT (Devlin, Chang, Lee, & Toutanova, 2019)), decoder (e.g., GPT-1 (Radford, Narasimhan, Salimans, & Sutskever, 2018) and GPT-2 (Radford, et al., 2019)), and encoder-decoder (e.g., BART (Lewis, et al., 2020)) models (Raiaan, et al., 2024; Minaee, et al., 2024), Advanced LLMs, such as GPT-3+ extend and combine such methods at such scale as to make seemingly intelligent agents (Zhao, et al., 2023).

2.1. Reasoning Abilities of Large Language Models (LLMs)

From a cognitive science perspective, reasoning is a set of mental processes where initial information is expanded upon through inferencing (Kurtz, Gentner, & Gunn, 1999). From the initial release of LLMs, researchers have been interested in how well LLMs can align with human reasoning (Wang, et al., 2023; Mitchell & Krakauer, 2023; Lampinen, et al., 2022). After evaluating how well or poorly an LLM compares to humans, researchers then focus on why it performed the given way, which leads to the evaluation of LLMs' sense of "logic" (Binz & Schultz, 2023).

The logical reasoning of LLMs has been extensively studied in a variety of capacities (Liu, et al., 2023; Huang & Chang, 2023; Luo, et al., 2024; Mondork & Plank, 2024; Lampinen, et al., 2022). There are many types of reasoning with the most popular types being inductive, deductive, and abductive (Luo, et al., 2024; Wang, et al., 2024). Deduction takes a top-down approach where a general statement is specified for an individual observation (Merriam Webster, 2024). Induction is a bottom-up approach where observations lead to a generalized conclusion (Merriam Webster, 2024). Abduction considers observations and derives the most reasonable conclusion (Merriam Webster, 2024). These types of reasoning are shown for example problems in Table 1. When using deduction, a correct conclusion is always guaranteed; however, that is not always the case when reasoning by induction and abduction.

Table 1. Types of Reasoning Examples

Deduction	1. All LLMs have pretraining data.
	2. GPT-4 is an LLM.
	3. GPT-4 has pretraining data
Induction (Correct Conclusion)	1. GPT-4 is an LLM.
	2. GPT-4 has pretraining data
	3. All LLMs must have pretraining data.
Induction (Incorrect Conclusion)	1. GPT-4 is an LLM.
	2. GPT-4 performs with 90% accuracy on Dataset A.
	3. All LLMs must perform with 90% accuracy with Dataset A.
Abduction (Correct Conclusion)	1. All LLMs have pretraining data.
	2. GPT-4 has pretraining data.
	3. GPT-4 must be an LLM.
Abduction (Incorrect Conclusion)	1. All LLMs have pretraining data.
	2. The artificial neural network (ANN) has pretraining data.
	3. The ANN must be an LLM.

The majority of LLM reasoning research has focused on logical reasoning as a whole rather than individual types. Although, some articles focus on a particular type such as deductive reasoning (Seals & Shalin, 2024) or strategic reasoning (Zhang, et al., 2024). One type of reasoning gaining in popularity is analogical reasoning, a type of abductive reasoning (Wang, et al., 2024).

2.2. Analogical Reasoning

Analogical reasoning utilizes analogies to infer information between a source (or familiar) scenario and a target (or unfamiliar) scenario (Kurtz, Gentner, & Gunn, 1999). Analogical reasoning has a relatively brief history starting with the ANALOGY algorithm (Evans, 1964), but not truly flourishing until the release of the Structure Mapping Engine (Gentner, 1983; Falkenhainer, Forbus, & Gentner, 1986) and the Analogical Constraint Mapping Engine (Holyoak & Thagard, 1989) in the late 1980s. The majority of early analogical reasoning work was led by cognitive scientists focused on verbal/text-based analogies (Combs K. , Bihl, Ganapathy, & Staples, 2022; Mitchell, 2021).

Analogical reasoning has been an effective way to infer information about an unfamiliar scenario in few-shot and zero-shot learning (Webb, Fu, Bihl, Holyoak, & Lu, 2023). However, it involves higher cognitive processes beyond the more simplistic processes such as information recall, which LLMs perform quite well on (Mitchell & Krakauer, 2023). Measuring the ability to which LLMs can learn by analogy or even create analogies, indicates more human-like reasoning abilities (Mitchell, 2021). Some researchers have tested this by LLMs’ abilities to generate analogies (Bhavya, Xiong, & Zhai, 2022; Jiayang, et al., 2023; Sultan, Bitton, Yosef, & Shahaf, 2024). Some researchers propose that LLMs have this ability based on performance on various testbeds (Webb, Holyoak, & Lu, 2023; Webb, Holyoak, & Lu, 2024); however, other researchers in the field disagree (Hodel & West, 2023; Lewis & Mitchell, 2024).

Text-based analogical reasoning has focused on four three types of analogies: word-based, sentence-based, and story-based/long-text (Ichien, Lu, & Holyoak, 2020; Combs, Lu, & Bihl, 2023). Word-based analogies come in the form, “A is to B as C is to D,” and are the simplest type of analogy. Sentence-based analogies are slightly more complex such as “She is growing like a weed.” Finally, the most complex analogies are long-text analogies, which involve stories with elements, plots, and settings that are compared and contrasted. We are interested in long-text analogies since they are the most complex test of LLMs’ reasoning

abilities. Considered to be long-text analogies, but consisting of sentences, the StoryAnalogy dataset was created to test the analogical reasoning abilities of LLMs (Jiayang, et al., 2023). Another dataset, AnaloBench, was designed such that an LLM could select the most analogous story from a story bank with varying lengths of each story and the number of stories in the bank (Ye, et al., 2024).

3. Methodology

This section describes the methodology applied to compare the various LLMs shown in Table 2. Nine open-source models and five proprietary models were selected for comparison due to popularity and ease of access. All of the open-source models were accessed via Ollama (<https://ollama.com>). The five proprietary models (GPT-3.5T, GPT-4, GPT-4o, Claude 3, and Gemini) were accessed via their web interface. Many models are identified by the number of parameters rather than a version name or number; however, if one existed it was included. All default hyperparameters were kept for the analysis which occurred in April 2024. The remainder of this section is presented as such. First, in Section 3.1, we present the long-text analogy datasets considered and ultimately selected for this analysis based on criteria determined by the authors. Then, in Section 3.2, we describe the prompt template and format used for all the LLMs in consideration.

Table 2. Selected LLMs for Analysis

Creator	Model	Ver.	Num. of Parameters
Google	Gemini	Pro	Unknown
	Gemma	1.1	7.8B
Meta AI	Llama 2		7B
	Llama 3		8B
Mistral AI	Mistral	0.2	7B
	Mixtral	8x	7B
Stability AI	StableLM 2		1.6B
	Stable Beluga		7B
Microsoft	Phi-3	Mini	3.8B
OpenAI	GPT-3.5T	Turbo-0125	175B
	GPT-4	Turbo	1.8T
	GPT-4o	2024-05-13	Unknown
Anthropic	Claude 3	Sonnet	Unknown

3.1. Long-text Analogy Datasets

Ichien, Lu, & Holyoak (2020) surveyed the literature for verbal analogy datasets and identified

seven story/long text datasets shown in Table 3, which we initially considered for our analysis. The entire testbed is available online (see data availability statement below). Modeling our analysis on the previous (Webb, Holyoak, & Lu, Emergent analogical reasoning in large language models, 2023) study, we are interested in retrieval problems (as classified by (Ichien, Lu, & Holyoak, 2020)), where an LLM is provided a source story and multiple options to choose from for the “most” analogous one. This criterion eliminates the Gentner, Gick, Gick2, and Keane datasets due to not being presented in story form. Additionally, the Clement dataset was excluded since the four problems were focused on the identification of similar aspects between stories rather than the selection of the most similar story given several options. From our initial dataset consideration shown in Table 3, we are left with the Rattermann and Wharton datasets.

Table 3. Long-text Analogy Datasets (modified from (Ichien, Lu, & Holyoak, 2020))

Dataset	Citation	Num. of Analogies
Gick	(Gick & Holyoak, 1980)	1
Gick2	(Gick & Holyoak, 1983)	2
Gentner	(Gentner & Toupin, 1986)	54
Keane	(Keane, 1987)	1
Clement	(Clement & Gentner, 1991)	4
Rattermann	(Gentner, Rattermann, & Forbus, 1993)	16*
Wharton	(Wharton, et al., 1994)	14

*Originally 18 analogies, but two were incomplete

The Rattermann dataset consists of 16 complete analogies (Sets #3 and #16 were incomplete) with a source story and five other stories (labeled A-E) with three similar (or dissimilar) aspects:

1. Entities – objects and characters,
2. First-order relations – relationship between an entity and a property or between entities,
3. Higher-order relations – more complex relationships that may occur between first-order relations or many-to-one relationships between entities/properties/relations.

Story A is a “literally similar story” because it mimics the source story’s entities and both sets of relationships. Story B is a “true analogy story” meaning that its entities are dissimilar; however, the relationships are similar to the source story. Story C is a “false analogy story” where its first-order relations match the source story but its entities and higher-order relationships do not. Story D is a “mere-appearance match” such that its entities and first-order relationships match, but not its higher-order relationships. Finally, Story E is a “new mere-appearance match” such that only the entities match the

source story. This is visually presented in Table 4 along with the rank of each story, as determined by the authors, from the most analogous (Story A) to the least analogous (Story E). Despite both Story B and C having two similar elements including first-order relations, Story B is believed to be more similar because higher-order relationships are maintained which requires more advanced cognitive parallels compared to similar entities, which Story D has. Similarly, Story C is believed to be more analogous to Story E because to draw parallels because the identification of first-order relations is a more difficult cognitive process compared to identifying similar entities. Note from this point on, the Rattermann results will be presented in order from most analogous to least analogous (as justified above): A, B, D, C, and E.

Table 4. Rattermann Dataset Structure

Story Elements	A	B	C	D	E
Entities	X			X	X
First-order relations	X	X	X	X	
Higher-order relations	X	X			
Most Analogous Order	1	2	4	3	5

The Wharton dataset consists of 14 analogies consisting of four stories: A1, B1, A2, and B2, where A1 is analogous with A2 and B1 is analogous with B2. The dataset is partially derived from previous studies in (Seifert, McKoon, Abelson, & Ratcliff, 1986; Rattermann & Gentner, 1987). Unlike in the Rattermann dataset, outside of the matching story, there is no rank order of the remaining two stories being “more” or “less” analogous with the source story. Each analogy is used twice, first with A1 serving as the source story with B1, A2, and B2 as options for the LLM to choose from. The second time, B1 serves as the source story and the LLM must choose between A1, A2, and B2 for being most analogous.

3.2. LLM Prompt Template & Format

Since the Rattermann and Wharton datasets are provided in different formats, the prompts used with the LLMs differ between the two. For the Rattermann data, the following prompt was used:

```
Consider the following story:
Source Story: <<Source Story>>
Now consider five more stories:
Story A: <<Story A>>
Story B: <<Story B>>
Story C: <<Story C>>
Story D: <<Story D>>
Story E: <<Story E>>
Rank Stories A, B, C, D, and E in order
from most analogous to least
analogous to the source story.
```

For the Wharton dataset, two prompts were used depending on the source story. When A1 was used as the source story the following prompt was used:

```
Consider the following story:
Source Story: <<Story A1>>
Now consider three more stories:
Story A: <<Story B1>>
Story B: <<Story A2>>
Story C: <<Story B2>>
Rank Stories A, B, and C in order from
most analogous to least analogous
to the source story.
```

The Wharton prompt was reused with B1 as the source story, A1 as Story A, A2 as Story B, and B2 as Story C. At no point in the study were the LLMs told whether their rankings were correct or incorrect. The story options were not randomized for two reasons. First, this made the evaluation easier. Second, the LLMs were unlikely to “learn” the ranking pattern since they were never informed of the “correct” answer to the prompt.

All open-source models, accessed via Ollama, were queried individually for each analogy problem set. However, the four proprietary models were queried in a continuous chat for each version of the datasets (one for Rattermann, one for Wharton with A1 as the source story, and one for Wharton with B1 as the source story).

4. Metrics

The metrics of interest look at LLM performance related to the selection of the correct story rankings and overall accuracy on the datasets. Since the Rattermann and Wharton datasets come in different formats, their metrics also differ.

For the Rattermann dataset, we consider five metrics. First, is overall rank accuracy, which is the percentage of the ranks the LLM assigned each story for each problem. This is a percentage measured out of 80, which would indicate that the LLM’s predicted rankings matched the actual rankings for each of the 5 stories in all 16 problems (note that this is measured out of 75 for Llama 2 and Gemma due to them providing invalid responses to the prompt). Other confusion matrix metrics were considered such as precision, recall, and F1-scores, but since we have an equal distribution of stories/ranks, the precision, recall, and F1-scores are all equivalent within a story/rank. However, we do consider the average predicted rank for each story. This is a value range from 1 (best, most analogous), to 5 (worst, least analogous). The “ideal” value for this metric depends on the Story, with each rank from low to high corresponding to Stories A, B, D, C, and E, respectively. The overall rank accuracy provides one number indicating how accurate the LLM was able to rank each story. However, the rank accuracy per story allows us to see which stories, and their corresponding

elements (entities, first-order relations, and higher-order relations), caused the LLM to rank one story higher or lower over another.

For the Wharton dataset, we consider two metrics. The first metric is accuracy, which is the percentage of the number of times the actual most analogous story (A2 if A1 is the source story and B2 if B1 is the source story), was correctly predicted by the LLM as the most analogous. This metric is measured as a percentage correct out of 28 total possible problems. The second metric is the average predicted rank of the actual most analogous story (again, A2 is A1 is the source story and B2 is B1 is the source story). This metric looks at the average rank of the actual most analogous story the LLM assigned for all 28 problems. This value ranges from 1 (best, considered most analogous) to 3 (worst, considered least analogous). These two metrics for the Wharton dataset are calculated for all the LLMs. Similar to the Rattermann metrics, the accuracy provides a percentage to describe the performance of the model; however, the average rank helps quantify the degree to which an LLM was incorrect when assigning its ranking.

5. Results & Discussion

The methodology discussed in Section 3 was applied to 13 models, and the metrics defined in Section 4 were applied. The results thereof are presented in Section 5.1 for the Rattermann dataset and Section 5.2 for the Wharton dataset.

5.1. Rattermann Dataset Results

The Rattermann dataset was evaluated on five metrics, overall rank accuracy and the average predicted rank per story. The overall rank accuracy is presented for all the LLMs in Table 5, which is the sum of their respective confusion matrices shown in Tables 5-17. The diagonals of the matrices are highlighted with a higher number on the diagonal being more ideal (shown in a darker color). The average predicted rank per story is shown for all the models in Table 19.

Table 5. GPT-4 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	16	0	0	0	0
2	0	4	12	0	0
3	0	12	4	0	0
4	0	0	0	16	0
5	0	0	0	0	16

Table 6. GPT-4o Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	16	0	0	0	0
2	0	8	8	0	0
3	0	7	4	5	0
4	0	1	4	11	0
5	0	0	0	0	16

Table 7. Gemini Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	15	0	1	0	0
2	1	2	13	0	0
3	0	13	1	2	0
4	0	1	1	14	0
5	0	0	0	0	16

Table 8. Claude 3 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	16	0	0	0	0
2	0	2	14	0	0
3	0	10	2	2	2
4	0	4	0	12	0
5	0	0	0	2	14

Table 9. Stable Beluga Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	12	0	1	0	3
2	2	11	0	3	0
3	0	3	3	10	0
4	1	1	12	2	0
5	1	1	0	1	13

Table 10. StableLM2 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	10	0	3	0	3
2	5	9	1	1	0
3	1	4	3	8	0
4	0	1	7	5	3
5	0	2	2	2	10

Table 11. Phi-3 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	16	0	0	0	0
2	0	2	12	2	0
3	0	2	4	8	2
4	0	8	0	3	5
5	0	4	0	3	9

Table 12. Mixtral Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	12	0	4	0	0
2	4	4	8	0	0
3	4	4	8	0	0
4	0	6	2	3	4
5	0	2	0	4	10

Table 13. Gemma Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	7	0	6	0	2
2	8	5	2	0	0
3	0	3	1	8	3
4	0	5	5	5	0
5	0	2	1	2	10

Table 14. Mistral Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	8	0	7	0	1
2	8	1	5	1	0
3	0	3	2	7	5
4	0	4	2	5	5
5	0	8	0	3	5

Table 15. Llama-3 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	10	0	6	0	0
2	6	2	7	0	1
3	0	3	2	5	6
4	0	7	1	4	4
5	0	4	0	7	5

Table 16. GPT-3.5T Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	4	0	13	0	0
2	13	1	3	0	0
3	0	4	0	4	8
4	0	6	0	7	3
5	0	6	0	5	5

Table 17. Llama-2 Confusion Matrix

	Stories				
Rank	A	B	D	C	E
1	4	0	0	0	11
2	4	3	7	0	1
3	0	1	2	12	0
4	0	10	4	1	0
5	7	1	2	2	3

In Table 18, the LLMs are presented in descending order from most accurate to least accurate. All accuracies were taken out of a total of 80 correct possible (16 sets multiplied by 5 stories each) except for two models which had a total of 75 correct possible. Set #17 had material that triggered Llama 2’s safety filter, causing it to refuse to answer and Gemma provided an invalid response to Set #6 such that only stories A and E were ranked. GPT-4 and GPT-4o performed higher than its competitors at 70% and 68.8%, respectively. The other models performed relatively close to one another when looking at overall rank accuracy.

Table 18. Rattermann Overall Rank Accuracy

Model	Number Correct	Accuracy
GPT-4	56	70%
GPT-4o	55	68.8%
Gemini	48	60%
Claude 3	46	57.5%
Stable Beluga	41	51.3%
StableLM2	37	46.3%
Phi-3 Mini	34	42.5%
Mixtral	31	38.8%
Gemma*	28	37.3%
Llama 3	23	28.8%
Mistral	21	26.3%
GPT-3.5T	15	18.8%
Llama 2*	13	17.3%

*Out of 75 instead of 80 due to invalid answer(s)

Next, we consider the average predicted rank per story. In a perfect scenario, Story A would all have rank 1 (darker color), Story B would have rank 2, and so on for the order of the columns in Table 19 (lighter color). The models are presented in the same order as they appear in Table 17 There is insufficient space to discuss all the results; however, it appears that the vast majority of LLMs performed relatively well on the rankings of Stories and A and E. Story D had a higher average rank across all the LLMs compared to Stories B and C, which suggests that similar story entities have an important impact on the rank determination. However, the presence of similar entities may distract LLMs from selecting the correct rankings for each story. For the bottom-performing models, the average rank for Stories B and C is close, which suggests that LLMs are unable to recognize when higher-order relations are present or not present in a given story. For more details regarding the story rankings of each model, see the confusion matrices in Tables 5-17.

Table 19. Rattermann Average Predicted Rank Per Story

Model	A	B	D	C	E
GPT-4	1	2.75	2.25	4	5
GPT-4o	1	2.56	2.75	3.69	5
Gemini	1.06	2.94	2.13	3.88	5
Claude 3	1	3.13	2.13	4	4.75
Stable Beluga	1.56	2.5	3.63	3.06	4.25
StableLM2	1.44	2.75	3.25	3.5	4.06
Phi-3 Mini	1	3.88	2.25	3.44	4.44
Mixtral	1.25	3.38	2.13	3.69	4.5
Gemma	1.53	3.27	2.53	3.6	4.07
Llama 3	1.38	3.81	1.88	4.13	3.81
Mistral	1.5	4.19	1.94	3.63	3.81
GPT-3.5T	1.81	4.13	1.19	4.06	3.81
Llama 2	3.13	3.6	3.07	3.33	1.87

5.2. Wharton Dataset Results

The Wharton dataset was evaluated on two metrics, overall accuracy and the average predicted rank for the actual most analogous story. The models are ordered from highest to lowest accuracy, and average rank was used to break a tie in accuracy in Table 20. Accuracy is reported out of 28 total possible problems for all the models except for StableLM2, which provided an invalid answer for Set #11 where A1 was the source story. By random chance, one should have an accuracy of 33.3%; however, many of the LLMs fall below this threshold. The highest accuracy of 46.4% was achieved by GPT-4o. The only other models to exceed random chance accuracy were GPT-3.5T, GPT-4, and Claude 3.

Table 20. LLM Results on Wharton Dataset

Model	Accuracy	Average Rank
GPT-4o	46.4%	1.929
GPT-3.5T	42.9%	1.857
GPT-4	42.9%	1.929
Claude 3	39.3%	2.036
Mistral	32.1%	2
Phi-3 Mini	32.1%	2
Llama 3	21.4%	2.179
Gemma	21.4%	2.25
Stable Beluga	17.9%	2.286
StableLM2	14.8%*	2.3
Llama 2	14.3%	2.214
Gemini	10.7%	2.429
Mixtral	0%	2.536

*Accuracy reported out of 27 total problems

The average rank is the average rank that the correct most analogous story achieved for the entire dataset for each model where a lower rank is ideal (meaning the model typically selected it as the most analogous story). The lowest average rank was achieved by GPT-3.5T, which was slightly less accurate than the best model, GPT-4o. This suggests that despite being the most accurate, when GPT-4o was wrong, it mis-ranked the most analogous story to being the least accuracy story more often than GPT-3.5T.

6. Conclusions

Over the past four years, large language models (LLMs) have had a significant impact on our world given their adept ability to receive, understand, and return information. However, sometimes this ability is inaccurately equated with reasoning, a higher-level cognitive process. To test this, we have benchmarked thirteen popular LLMs on two analogical reasoning datasets named Rattermann (Gentner, Rattermann, & Forbus, 1993) and Wharton (Wharton, et al., 1994). Tasked with ranking stories from most to least analogous to a source story, the LLMs had varying performances on both datasets. For the Rattermann dataset, the high-performing LLMs obtained accuracies reaching up to 70%. However, the LLMs' rankings indicate that similar story entities (character(s) and setting(s)) can distract LLMs from other aspects of the stories such as first-order and higher-order relations. Additionally, LLMs may not be able to realize the presence of higher-order relationship(s) within a given pair of stories. The top 3 LLMs on the Rattermann dataset were GPT-4, GPT-4o, and Gemini, respectively. The Wharton dataset was more difficult for the LLMs, the highest accuracy was only 46.4% and nine LLMs had an accuracy lower than random chance (33%). The top 3 LLMs on the Wharton dataset were the GPT-family of algorithms GPT-4o, GPT-3.5T, and GPT-4. Overall, we notice that proprietary models have an advantage on higher-level cognitive tasks such as analogical reasoning and identify it as an area of improvement for future LLM development. Future research could look at additional LLMs and analogy datasets as well as a deeper analysis of the results. A potential follow-on study for the conversational models (ChatGPT, Gemini, Claude, etc.) is to tell them the correct answer and to analyze the reasoning behind their rankings.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily represent any views of

the U.S. Government, U.S. Department of Defense, or U.S. Air Force. This work was cleared for unlimited release under AFRL-2024-2924. This research was partially funded by the Air Force Research Laboratory through the Sensing, Learning, Autonomy, and Knowledge Engineering (SLAKE) contract (FA8650-19-C-1692).

Data Availability

The datasets used are available online in the "Cognitive Psychology.xlsx" file found at cvl.psych.ucla.edu/resources/AnalogyInventory.zip.

References

- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessment: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Bhavya, B., Xiong, J., & Zhai, C. (2022). Analogy generation by prompting large language models: A case study of InstructGPT. *Proceedings of the 15th International Conference on Natural Language Generation* (pp. 298-312). Waterville: ACL.
- Binz, M., & Schultz, E. (2023). Turning large language models into cognitive models. *arXiv:2306.03917*.
- Bommineni, V. L., Bhagwagar, S., Balcarcel, D., Bommineni, V., Davazitkos, C., & Boyer, D. (2023). Performance of ChatGPT on the MCAT: The road to personalized and equitable premedical learning. *MedRxiv*, 1-19.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*. Virtual.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., . . . Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 39.
- Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89-132.
- Combs, K., Bihl, T. J., & Ganapathy, S. (2024). Utilization of generative AI for the characterization and identification of visual unknowns. *Natural Language Processing Journal*, 7, 100064.
- Combs, K., Bihl, T. J., Ganapathy, S., & Staples, D. (2022). Analogical reasoning: An algorithm comparison for natural language processing. *Proceedings of the 55th Hawaii International Conference on System Sciences* (pp. 1310-1319). HICSS.
- Combs, K., Lu, H., & Bihl, T. J. (2023). Transfer learning and analogical inference: A critical comparison of algorithms, methods, and applications. *Algorithms*, 146.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis.

- Dodgson, N. (2023). Artificial intelligence: ChatGPT and human gullibility. *Policy Quarterly*, 19(3), 19-24.
- Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. *Proceedings of the April 21-23, 1964, spring joint computer conference*. New York City.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). The structure-mapping engine. *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence* (pp. 272-277). Philadelphia: AAAI.
- Gemma Team, Google DeepMind. (2024). Gemma: Open models based on Gemini research and technology. *arXiv:2403.08295*, 1-17.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 10(3), 277-300.
- Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.
- Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25, 524-575.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306-355.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1-38.
- Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.
- Google. (2023, May 10). *PaLM 2 Technical Report*. Google. Retrieved from Google AI: <https://ai.google/discover/palm2>
- Hadi, M. U., Al Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Muhammad, I., . . . Mirjalili, S. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1-45.
- Hodel, D., & West, J. (2023). Response: Emergent analogical reasoning in large language models. *arXiv*, arXiv:2308.16118.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive science*, 295-355.
- Huang, J., & Chang, K. C.-c. (2023). Toward reasoning in large language models: A survey. *arXiv:2212.10403*.
- Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior research methods*, 52(5), 1803-1816.
- Jiayang, C., Qiu, L., Ho, C. T., Fang, T., Wang, W., Chan, C., . . . Zhang, Z. (2023). StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv:2310.12874*.
- Katz, D. M., Bommarito, M. J., Gao, S., & Pablo, A. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society*, 382, 20230254.
- Keane, M. (1987). On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology Section A*, 39(1), 29-41.
- Kurtz, K. J., Gentner, D., & Gunn, V. (1999). Reasoning. In B. M. Bly, & D. E. Rumelhart, *Cognitive Science* (pp. 145-200). Academic Press.
- Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., . . . Hill, F. (2022). Language models show human-like content effects on reasoning tasks. *arXiv:2207.07051*.
- Lewis, M., & Mitchell, M. (2024). Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv:2402.08955*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871-7880). Online: ACM.
- Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv*, arXiv:2304.03439.
- Liusie, A., Manakul, P., & Gales, M. (2024). LLM comparison assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 139-151). Malta: EACL.
- Luo, M., Kumbhar, S., Shen, M., Parmar, M., Varshney, N., & Baral, C. (2024). Toward LogiGLUE: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *arXiv:2310.00836*.
- Merriam Webster. (2024). *Deduction vs Induction vs Abduction*. Retrieved from Merriam-Webster Dictionary: <https://www.merriam-webster.com/grammar/deduction-vs-induction-vs-abduction>
- Meta. (2023, July 18). *Meta and Microsoft introduce next generation of llama*. Retrieved from Meta Newsroom: <https://about.fb.com/news/2023/07/llama-2/>
- Meta. (2024, April 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. Retrieved from Meta Blog: <https://ai.meta.com/blog/meta-llama-3/>
- Meta AI. (2023, February 24). *Introducing LLaMA: A foundational, 65-billion-parameter large language model*. Retrieved from Meta AI Blog: <https://ai.facebook.com/blog/large-language-model-llama-meta-ai/>
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv:2402.06196*, 1-43.
- Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1), 79-101.
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.
- Mondork, P., & Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey. *arXiv:2404.01869*.
- Narang, S., & Chowdhery, A. (2022, April 4). *Pathways Language Model (PaLM): Scaling to 540 billion parameters for breakthrough performance*. Retrieved from Google Research Blog: <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

- Niesler, T., & Woodland, P. (1996). A variable-length category-based n-gram language model. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 164-167.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI White Paper*. Retrieved from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *1, 8*. OpenAI.
- Raiaan, M. A., Mukta, M. S., Fatema, K., Fahad, N. M., Sakib, Z., Mim, M. M., . . . Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.
- Rattermann, M. J., & Gentner, D. (1987). Analogy and similarity: Determinants of accessibility and inferential soundness. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 23-25). Amhurst: Cognitive Science Society.
- Seals, S. M., & Shalin, V. L. (2024). Evaluating the deductive competence of large language models. *arXiv:2309.05452*.
- Seifert, C. M., McKoon, G., Abelson, R. P., & Ratcliff, R. (1986). Memory connections between thematically similar episodes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(2), 220-231.
- Sultan, O., Bitton, Y., Yosef, R., & Shahaf, D. (2024). ParallelPARC: A scalable pipeline for generating natural-language analogies. *arXiv:2403.01139*.
- Wang, Y., Chen, W., Xiaotian, H., Lin, X., Zhao, H., Liu, Y., . . . Yang, H. (2024). Exploring the reasoning abilities of multimodel large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv:2401.06805*.
- Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Wuang, W., . . . Liu, Q. (2023). Aligning large language models with humans: A survey. *arXiv:2307.12966*.
- Webb, T., Fu, S., Bihl, T., Holyoak, K. J., & Lu, H. (2023). Zero-shot visual reasoning through probabilistic analogical mapping. *Nature Communications*, *14*(1), 5144.
- Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526-1541.
- Webb, T., Holyoak, K. J., & Lu, H. (2024). Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *arXiv:2404.13070*.
- Wharton, C. M., Holyoak, K. J., Downing, P. E., Lange, T. E., Wickens, T. D., & Melz, E. R. (1994). Below the surface: Analogical similarity and retrieval competition in reminding. *Cognitive Psychology*, *26*, 64-101.
- Ye, X., Wang, A., Choi, J., Lu, Y., Sharma, C. S., Lingfeng, . . . Andrews, N. K. (2024). AnaloBench: Benchmarking the identification of abstract and long-context analogies. *arXiv:2402.12370*.
- Zhang, T., Mao, S., Ge, T., Wang, X., de, W. A., Xia, Y., . . . Wei, F. (2024). LLM as a mastermind: A survey of strategic reasoning with large language models. *arXiv:2404.01230*.
- Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., . . . Du, Y. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.