

Cognitive, Behavioral, and Social Data

Analogical Reasoning with Llama 2



Department of Psychology
Università degli Studi di Padova

Fairouz Baz Radwan, Nour Al Housseini, and Sofia Pope Trogu

Supervisor: Prof. Giuseppe Sartori

February 9th, 2024

Table of Contents

1	Abstract	1
2	Introduction	1
2.1	Analogical Reasoning and debate on AI’s Intelligence	1
2.2	Reference Paper	2
2.2.1	Summary	2
2.2.2	Baselines	3
2.2.3	Models Studied in Paper	3
2.3	Llama 2	4
2.4	Final Overview	4
3	Materials and Methods	4
3.1	Dataset	4
3.2	Subsetting	5
3.3	MaxDiff Formalization and Prompt	6
3.4	Reversals	7
3.5	Performance Metrics	7
4	Results	8
4.1	Observations	8
4.2	Tables and Plots	8
5	Discussion	10
6	Conclusion	12
7	Appendix	13
7.1	A Taxonomy of Semantic Relations	13
8	References	17

1 Abstract

This study explores the intricacies of analogical reasoning, a fundamental cognitive process interwoven in the fabric of human cognition. Our investigation is carried out through the implementation of the "SemEval-2012 Task 2: Measuring Degrees of Relational Similarity" reliable framework. Through leveraging Llama 2, we aimed to assess its ability in discerning subtle variations within semantic relationships measured against established baselines. The analysis was followed with rigorous metrics, for instance employing Spearman's correlation and MaxDiff scores scrutinizing Llama's proficiency in identifying prototypical word pairs across a spectrum of semantic relations. Our findings were able to unveil relational nuances in Llama 2's responses, including an intriguing preference for extremal pairs and struggles with reversals. Even though the model was not able to replicate the performances of the best established models, its capacity for intricate relational reasoning highlights its potential in analogical reasoning contributions.

2 Introduction

2.1 Analogical Reasoning and debate on AI's Intelligence

In order to explain analogical reasoning, the concept of analogy should be introduced. Analogy has been recognized as playing an important heuristic role in human cognitive function, particularly aiding in intellectual discovery. It has been employed in a wide variety of settings to generate insight and formulate possible solutions to problems. This brings us to analogical reasoning, a built-in, automatic, cognitive inductive reasoning process that aids us in using information from one context to draw inferences or make predictions about another context. Through recognizing similarities between different situations of ranging understand-ability, we are able to apply prior knowledge and make connections under new contexts. This process is fundamental to human thinking and problem-solving. By leveraging knowledge from the familiar (source domain) to the less familiar (target domain), individuals are able to identify common similarities or differences to create more concrete, contextualized ideas. For example, if someone is trying to understand the concept of an atom and has prior knowledge of a solar system, they might use an analogy that likens electrons orbiting an atomic nucleus to planets orbiting the sun. This analogical connection can thus facilitate a better understanding of the atomic structure.

As previously mentioned, analogical reasoning is a cognitive process embedded in the structure of human awareness and recognized as a fundamental aspect of human thought. The flexibility in thinking serves as a cognitive foundation connecting various skills and processes into a unified intelligence. In addition to the mind's flexibility, another main characteristic of human intelligence is built upon the multifaceted nature of the human mind. Humans are rational beings able to solve a wide range of complex problems supplemented by predefined rules of logic, analytics, and ethics, which makes them the most intelligent species ever observed. This innate rationale present in humans defines "Real or Biological Intelligence" that uniquely sets apart humans from machines. However, humans tend to be limited in some mental capacities, such as ingrained cognitive biases and limited cognitive

capacities. As humans continue to push the boundaries of intelligence and survival, the landscape of artificial intelligent systems is evolving in parallel creating a dynamic interplay between the two. With regard to artificial intelligent systems, it is important to keep in mind some crucial considerations. Intelligent systems are not meant to be derogatory or demeaning, but rather the complete opposite. Artificial systems have been developed in a way to complement and promote human intelligence, particularly where it has shown lacking or weakness. The ultimate goal here is to allude to human intelligence and reach an end focus for AI. Regardless of one's benefit over the other, cultivating a series of cognitive tasks that illuminate human's limitations will foster the advancement of AI abilities.

The conceptualization of intelligence has become the target of many researchers in the field of mechanical development, artificial humanoid robots, as well as AI systems. Implementing an analogical reasoning task is an important area of cognitive research to pursue and has been the target of several studies, one of which will be addressed in the upcoming discussion.

2.2 Reference Paper

Being an expansive cognitive topic, analogical reasoning serves as a generalized term encompassing several intellectual functions, including relational similarity. This particular function represents an integral component of analogy in linking between source and target domains.

2.2.1 Summary

The article SemEval-2012 Task 2: Measuring Degrees of Relational Similarity offers an interesting implementation of the analogical reasoning task described. The researchers performed an experiment, in which they focused on the degree of relational similarity among instances, or word pairs in a relation class (word X:word Y). The greater the degree of similarity between instances, the greater the likelihood that knowledge will be transferred from the source to the target domain. This will generate a high relational similarity between the instances and they will thus be labeled as analogous.

This paper introduces a SemEval task that differs from previous tasks by measuring the degree of prototypicality for instances within a given relation class, rather than just classifying relations into discrete categories. Prototypicality can be defined as how representative or illustrative an instance is to a given relation class. The paper presents the first dataset of calculated relational similarity ratings across 79 different relation subcategories, which allows for a more nuanced understanding of relational similarity. The 79 relation categories represent a hierarchical system that are clustered into ten categories. This approach captures the variability within a class of analogous relations, which was not addressed in previous SemEval tasks that focused largely on discrete classification without considering continuous degrees of similarity. The task evaluated by the researchers was split into two important phases. Phase One focused on data collection of new examples of word pairs for each subcategory of the semantic relations. Phase Two involved having the Turkers select the most and least illustrative pairs from pairs generated in phase one, or their degree of prototypicality within the specified relation classes. Prototypicality is then measured using

a MaxDiff approach where a word pair is scored based on the percentage of times it is chosen as the most illustrative minus the percentage of times it is chosen as the least illustrative.

2.2.2 Baselines

The paper discusses two baselines used for further evaluating the prototypicality of word pairs within semantic relation classes: Random and PMI (Pointwise Mutual Information). The term baseline refers to a standard or reference point against which the systems’ performance can be compared to, thus evaluating their ability to rate the prototypicality of word pairs. The Random baseline assigns ratings randomly to each pair in a subcategory, with an expected Spearman correlation of zero and an expected MaxDiff score of around 31% due to instances where pairs receive equal votes from the Turkers. The PMI baseline, on the other hand, rates pairs based on the statistical association between the words, using a PMI score calculated from a large corpus of approximately 50 billion tokens. The PMI baseline is more sophisticated than the Random baseline as it considers the actual linguistic association between words. Baselines are important because they provide a means to assess whether the systems are performing better than simple, non-informative strategies such as random guessing or basic statistical association measures like PMI. The effectiveness of the systems is measured against these baselines to determine if they can significantly outperform these basic approaches.

2.2.3 Models Studied in Paper

The authors of the paper wanted to provide an exhaustive study on the relational similarity task, so they asked three academic institutions to test the SemEval task on their artificial systems. The teams are from Benemérita Universidad Autónoma de Puebla (BUAP), University of Texas at Dallas (UTD), and University of Minnesota, Duluth (Duluth). From these three institutions, six different models were used. Each system offered a new approach and is briefly described:

BUAP’s system uses multiple features including lexical, intervening words, WordNet relations, and syntactic features to represent each pair as a vector. Prototypicality is based on cosine similarity with the relation class’s pairs. UTD’s system has two approaches: NB and SVM. In UTD’s system, NB utilizes unsupervised learning to identify patterns between word pairs, and it assigns a ranking to each pattern based on how detailed or specific it is within its subcategory. SVM finds patterns similarly but uses them as feature vectors for an SVM classifier, with prototypicality ratings based on SVM confidence. Duluth’s system has three versions (V0, V1, V2), all using WordNet to build sets of concepts connected to the pairs’ words. The distinction between each version is how much it expands to encompass related concepts, and we assess the typicality by measuring the similarity of the definitions or explanations of these concepts when they are combined into a single text, referred to as concatenated glosses. Table X reports the average Spearman’s correlation (ρ) and MaxDiff scores for all systems across test subcategories, as well as the number of subcategories where each system achieved statistically significant Spearman’s ρ . The systems’ performance varied across different subcategories, with UTD-NB performing above the PMI baseline in many cases. However, no single system achieved superior performance on all subcategories, indi-

cating the task’s complexity. The diversity of methodologies is crucial for the research task of rating degrees of prototypicality. Not only that, but they offer potential for improving applications that require nuanced distinctions between instances of the same relation.

2.3 Llama 2

Llama 2 is the chosen open-source large language model for the following research project. It is interesting to note that Llama 2 is not a single model, but rather a collection of three with an increasing population of parameters used to train the models (7B, 13B, 70B). The parameters’ richness stems from the necessity of optimizing model performance to ensure an effective management and optimization of the various stages involved in processing input, generating responses, and handling outputs. Being Meta’s equivalent of OpenAI’s GPT-4, Google’s PaLM 2, and Anthropic’s Claude 2, it has been trained on a wide variety of publicly available data accounting for a 40% increase in training data compared to its predecessor Llama 1. Llama 2 has also been trained on two trillion “tokens” accessible from Wikipedia, Common Crawl, and Project Gutenberg to create its architecture. Every token represents a word/phrase that empowers the model to define different texts as human-like as possible. However, it has shown mediocre performance when compared with the previously mentioned models. This stems from the fact that Llama 2 is a foundation model and not a fine-tuned one, meaning that it was built with the promise of future implementations and adaptations for improvement. The accessibility of Llama 2 is wide, as you can access an online AI platform on Quora’s Poe alongside a wide selection of other model chat interfaces. One can only access Llama 2 through the official Llama on Hugging Face that provides chatbots of the previous Llama models as well.

2.4 Final Overview

In summary, this paper will describe an implementation of the SemEval Task 2 introduced by the article, SemEval-2012 Task 2: Measuring Degrees of Relational Similarity, using Llama 2 70B, reporting and analysis on results, and final comparison to the other benchmark systems. We predict that Llama 2 will perform on par, but likely not better than the model systems presented in the reference paper given the LLM’s previously mentioned strengths and limitations.

3 Materials and Methods

3.1 Dataset

The dataset utilized in our study was constructed using surveys carried out by the paper’s authors on human participants. As previously explained, the dataset development involved a two-phase approach with contributions from Amazon Mechanical Turk workers. Initially, workers were presented with examples of word pairs representing a specific semantic relation and asked to generate additional pairs fitting the relation. Subsequently, the second phase involved assessing the similarity of these generated pairs to paradigm examples, ensuring a

nuanced understanding of relational similarity. A system of "check" questions was integral to this process, filtering data based on workers' comprehension and adherence to the relation's intended meaning, ensuring the dataset's quality and relevance. The dataset was composed of ten main relational categories further divided into 79 subcategories, with ten allocated for training and 69 for testing. The effort of the Turkers yielded an extensive collection of over 6,000 word pairs. In the following list, we provide a brief characterization of the 10 categories:

1. **CLASS INCLUSION:** one word names a class that includes the entity named by the other word.
2. **PART-WHOLE:** one word names a part of the entity named by the other word, or something that is characteristically not a part.
3. **SIMILAR:** one word represents a different degree or form of the object, action, or quality represented by the other word.
4. **CONTRAST:** one word names an opposite or incompatible of the other word.
5. **ATTRIBUTE:** one word names a characteristic quality, property, or action of the entity named by the other word.
6. **NON-ATTRIBUTE:** one word names a quality, property, or action that is characteristically not an attribute of the entity named by the other word.
7. **CASE RELATION:** one word names an action that the entity named by the other word is usually involved in, or both words name entities that are normally involved in the same action in different ways, e.g., as agent, object, recipient, or instrument of the action.
8. **CAUSE-PURPOSE:** one word represents the cause, purpose, or goal of the entity named by the other word, or the purpose or goal of using the entity named by the other word.
9. **SPACE-TIME:** one word names a thing or action that is associated with a particular location or time named by the other word.
10. **REPRESENTATION:** one word names something that is an expression or representation of, or a plan or design for, or provides information about, the entity named by the other word.

3.2 Subsetting

To maintain a focused and manageable scope within our study, we strategically selected a subset of this dataset. Our analysis concentrated on four main categories highlighted in the table above, encompassing 25 subcategories in the testing set. This subsetting allowed for a detailed yet efficient examination of relational semantics and model performance.

Relational Categories	Testing Set	Training Set
CLASS-INCLUSION	4	1
PART-WHOLE	8	1
SIMILAR	6	1
CONTRAST	7	1
ATTRIBUTE	7	1
NON-ATTRIBUTE	8	1
CASE RELATIONS	7	1
CAUSE-PURPOSE	8	1
SPACE-TIME	9	1
REFERENCE	5	1

Table 1: Number of subcategories per relational category for train-test split

For each subcategory, we dealt with approximately 100 unique sets of four-word pairs. To ensure optimal performance, we presented our language model with one set of word pairs at a time. Initially, we experimented with multiple groups simultaneously, but this approach adversely affected the model’s responses. After numerous trials, we settled on the more time-intensive method of single-set inputs, which yielded better results. Additionally, we crafted various versions of the MaxDiff questions using different prompts, ultimately selecting the most effective one through rigorous testing and refinement.

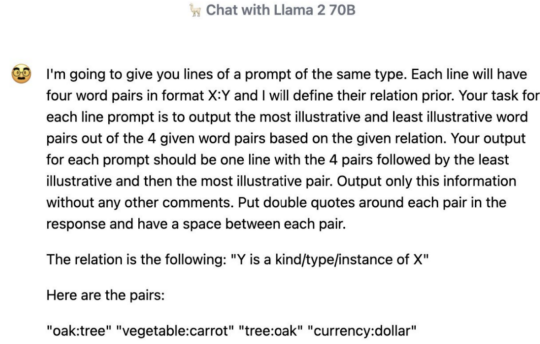
3.3 MaxDiff Formalization and Prompt

For the MaxDiff question replicated from the paper and originally presented to Turkers, we adapted a similar query for our language model. We provided a concise explanation of the task: presenting a specific relation along with four word pairs generated by the Turkers, and instructed the model to identify the pair that most exemplifies the given relation, as well as the pair that is least illustrative of it. This approach aimed to closely emulate the human input and response, as well as refine the model’s relational understanding.

In our quest to optimize the language model’s performance, we meticulously experimented with various prompt structures. The utilized technique is known as “prompt engineering”. This involved altering the wording, adjusting the level of detail, and observing the model’s response to these changes.

Our final selection of the most effective prompt was based on its ability to clearly communicate the task and the specified relation to the language model. The chosen prompt succinctly explained the task at hand and then directed the language model to identify the word pairs that were most and least illustrative of the given relation, ensuring a focused and

precise assessment of the model’s comprehension.



Chat with Llama 2 70B

😊 I'm going to give you lines of a prompt of the same type. Each line will have four word pairs in format X:Y and I will define their relation prior. Your task for each line prompt is to output the most illustrative and least illustrative word pairs out of the 4 given word pairs based on the given relation. Your output for each prompt should be one line with the 4 pairs followed by the least illustrative and then the most illustrative pair. Output only this information without any other comments. Put double quotes around each pair in the response and have a space between each pair.

The relation is the following: "Y is a kind/type/instance of X"

Here are the pairs:

"oak:tree" "vegetable:carrot" "tree:oak" "currency:dollar"

Figure 1: Example prompt provided to Llama 2

3.4 Reversals

One interesting aspect of the dataset was the incorporation of reversals, which served as a strategic method to assess the discernment capabilities of our Language Model. This inclusion was deliberate, aiming to challenge the LLMs in distinguishing and selecting the most and least representative word pairs for a given relation. By embedding these reversals, researchers could scrutinize the LLMs’ depth of understanding and their ability to navigate complex relational dynamics, providing a more robust evaluation of their semantic comprehension. An example of such word pairs in subsection 1a: *tree:oak* \rightarrow *oak:tree*.

This approach not only tested Llama’s ability to discern the most and least illustrative word pairs but also provided insight into its grasp of the relational context. A pattern of selecting both reversals as either the most or least illustrative could indicate a superficial understanding of the task, while consistently avoiding these pairs might suggest a deeper, more nuanced comprehension.

3.5 Performance Metrics

After gathering the responses from our language model, Llama 2, we employed the Perl code files accompanying the paper to compute various performance metrics. This step was crucial for assessing our model’s capabilities against the benchmarks set in the study, as well as against the gold standard ratings derived from Turkers’ responses. The MaxDiff score, a key metric, was meticulously calculated for each subcategory using the provided `score_maxdiff.pl` script, with results systematically documented in text files for straightforward analysis. This score, reflecting the accuracy of the model’s responses in alignment with the majority of Turkers, served as a primary indicator of performance.

In addition to the MaxDiff score, we also computed Spearman correlations for each subcategory, aiming to measure the degree of alignment between our model’s ranking of word pairs and the rankings produced by the Turkers. We utilized the `maxdiff_to_scale.pl` script, which processed Llama 2’s responses and scaled the word pairs. Subsequently, we used the `score_scale.pl` script, which compared these scaled Llama responses against the

Turkers’ scaled values. This meticulous process allowed for a precise evaluation of correlation between the model’s relational reasoning and human judgment, offering insightful metrics on the model’s performance in analogical tasks.

4 Results

4.1 Observations

To discern patterns in Llama 2’s responses to the MaxDiff questions, we devised a methodical analysis. We calculated the frequency of instances where Llama 2 selected the first or last pair in the list as the most or least illustrative of the given relation, Table 2. This evaluation was conducted across a random selection of subcategories from the 25 we focused on. Our findings highlighted a significant trend: Llama 2 often chose the pairs at the extremities of the list, particularly favoring the first pair as the most illustrative with notable frequency (exceeding 60% in numerous subcategories). Additionally, there was a distinguishable pattern of selecting the last pair as the least illustrative. Although less prevalent, we also observed instances where Llama 2 chose reversed pairs as the most and least illustrative, hinting at nuanced patterns in the model’s decision-making process.

In cases where the word pair lists contained inappropriate or sensitive language, Llama 2 sometimes opted not to provide a response. This reaction necessitated the exclusion of such word pairs from our data collection and analysis to maintain the integrity and appropriateness of our dataset. For example, the subcategory 3d: SIMILAR- Dimensional Naughty contained pairs like “f**k : r**e” which were removed from our study to ensure that our findings and the language model’s performance were evaluated within a suitable context.

4.2 Tables and Plots

Subcategory	Last pair as least illustrative	First pair as least illustrative	Last pair as most illustrative	First pair as most illustrative
1b	42.72	20.39	5.83	65.05
1c	6.67	87.62	25.71	9.52
2b	1.82	92.73	31.82	1.82
2f	25.00	2.78	10.19	75.93
2i	34.69	3.06	11.22	63.27
3d	24.21	3.16	9.47	71.58
3g	25.26	1.05	22.11	61.05
4d	65.33	1.33	1.33	98.67
4h	33.33	10.19	13.89	47.22

Table 2: Pattern analysis to understand the percentage of times Llama 2 chooses a particular pair as either the least or the most illustrative for example subcategories

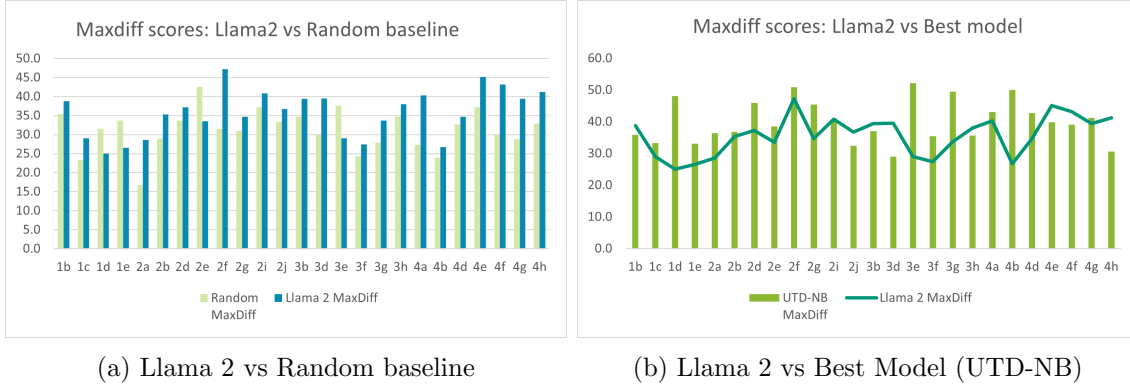


Figure 2: Comparison of Max Diff Scores between Llama 2 and Random Baseline (a) and between Llama 2 and the best model, UTD-NB (b).



Figure 3: Comparison of average Max Diff Scores computed over all subcategories in a given overall relation class and among all included systems in testing. Results presented for relation classes, CLASS-INCLUSION (top left), CONTRAST (top right), PART-WHOLE (bottom left), and SIMILAR (bottom right).

Team	System	Spearman’s ρ	# Subcategories $p < 0.05$	# Subcategories $p < 0.01$	Max Diff
BUAP	BUAP	0.014	2	0	31.70
UTD	NB	0.229	23	18	39.43
	SVM	0.116	11	5	34.68
Duluth	V0	0.0450	9	3	32.37
	V1	0.0387	10	4	31.47
	V2	0.0380	9	3	31.09
Baseline	Random	0.018	6	0	31.15
Meta	Llama 2	0.177	5	3	35.43

Table 3: Average Spearman’s ρ and MaxDiff scores for all system across all 69 test subcategories. Columns 4 and 5 denote the number of subcategories with a Spearman’s ρ that is statistically significant at the noted level of confidence. Note Llama 2 doesn’t include all testing subcategories

Relation Class	Class-Inclusion	Part-Whole	Similar	Contrast
Random	0.057	0.012	0.026	-0.049
BUAP	0.064	0.066	-0.036	0
UTD-NB	0.233	0.252	0.214	0.206
UTD-SVM	0.093	0.142	0.131	0.162
UMD-V0	0.045	-0.061	0.183	0.142
UMD-V1	0.178	-0.084	0.208	0.12
UMD-V2	0.168	-0.054	0.198	0.051
LLama 2	-0.016	0.241	0.157	0.307

Table 4: Average Spearman’s ρ correlation with the Turker rankings in each of the high-level relation categories, with the highest average correlation for each category shown in bold.

5 Discussion

Beyond our initial qualitative observations, we can evaluate Llama 2’s performance in the SemEval Task 2 on analogical reasoning quantitatively as well. As introduced in the methods section, two primary performance metrics can be utilized to rate the model: Max Diff Score and Spearman’s ρ . While the MaxDiff score reflects the accuracy of the model’s response compared to the response of the majority of Turkers, Spearman’s ρ is a measure of alignment in the prototypicality rating between each model and the Turker’s responses.

The first couple of figures illuminate some patterns for the MaxDiff Score. Figure 2 plots a comparison of the MaxDiff Scores for each subcategory on the x-axis between Llama 2 and

two benchmark models: Random baseline and UTD-NB. Compared to the Random Baseline, Llama 2 provides more accurate response in each subcategory, with a few exceptions: 1c, 1e, 2e, 3e. These subcategories represent Singular Collective, i.e. cutlery:spoon, ClassIndividual, i.e. queen:Elizabeth, PART-WHOLE, i.e. planting:gardening, and Conversion, i.e. apprentice:master, respectively. This may indicate that Llama 2 and other LLMs may have a difficult time in performing relational similarity tasks in these subcategories, especially if a random guess performs better than the "intelligent" system. We also compared Llama 2's performance with the best benchmark model, UTD-NB. On this granular subcategory level, it's difficult to comment on a general trend. However, UTD-NB often exhibits higher MaxDiff scores than Llama 2, particularly higher in some select subcategories: 1d, 2d, 3e, 3g. This might indicate that UTD-NB can better discern Plural Collective, i.e., groceries:eggs, Event:Feature, i.e., rodeo:cowboy, Conversion, i.e. apprentice:master, and Coordinates, i.e. son:daughter, respectively than Llama 2. Meanwhile, Llama 2 performs better in the following categories: 2j, 3b, 3d, 4e, 4f, and 4h, suggesting that advancements in LLMs with Llama 2 have enabled a better capturing of these relational similarities.

Figure 3, on the other hand, presents a comparison of average Max Diff scores for each tested relational category across all systems. Among all relation classes, UTD-NB remains the top-performing model. In CLASS-INCLUSION, Llama 2 ranks as the second lowest performing system; in CONTRAST, it holds the third-best position; in PART-WHOLE, it stands as the second-best system; and in SIMILAR, it ranks as the third worst system. The performance of Llama 2 across various relation classes lacks a discernible trend, with instances of comparable performance to the top model and other instances of significantly poorer performance.

The subsequent tables in the results section were an adaptation of tables presented in the reference paper and focus mainly on the Spearman correlation across all systems. Table 3 displays the average Spearman correlation and Max Diff scores by system. While the ρ and MaxDiff represent an average across all 69 test categories for the systems presented in the paper, the number of significant subcategories ($p < 0.05$ or $p < 0.01$) is based solely on the subcategories we tested on the Llama 2 model. We used this strategy in order to align the results from the paper with our results from Llama 2 determined by the subset of subcategories. Overall, the average Max Diff score for Llama 2 is the second highest among the systems, suggesting high performance, but still not as good as the best model, UTD-NB. The MaxDiff of Llama might be a bit skewed though, as the average is computed over fewer sub-categories than the other systems. The Spearman correlation values can be challenging to analyze on their own, but can be compared between different systems. During these comparisons, we can note that Llama 2 performs relatively well in terms of raw correlation value compared to the best model, UTD-NB. However, after running a one sample t-test, we found much fewer subcategories with significant ρ s than the best model.

Table 4 expands upon the previous evaluation of Spearman by listing the average correlation value in each aggregate relation category across all analyzed systems. The results indicate that UTD-NB continues to outperform all other systems in terms of this performance metric overall. However, Llama 2 successfully beats the UTD-NB system in the CONTRAST relation class, indicating Llama's potentially strong ability to identify contrasting relationships. The Llama 2 model also performs almost as well as the best model in the Part-Whole and Similar categories, yet does noticeably worse in Class-Inclusion. This last finding aligns

with the results we found for the Max Diff scores, in which Llama 2 performs poorly in subcategories included in Class-Inclusion particularly.

There may be several explanations for Llama’s varying performance on these analogical reasoning tasks. Llama 2 faces limitations in these tasks due to challenges in capturing subtle relational nuances and adapting to specific tasks without fine-tuning. Its generic training may hinder its ability to generalize effectively to diverse analogical reasoning scenarios, necessitating fine-tuning for optimal performance. Despite recognizing the potential benefits of fine-tuning for improving performance, the lack of sufficient data or necessary resources prevented its implementation to make a significant impact. Additionally, Llama 2 may struggle to learn from limited or noisy data, impacting its ability to discern underlying patterns and relationships in analogical reasoning tasks. Addressing these limitations could involve refining the model’s architecture, incorporating task-specific training data, and enhancing its ability to generalize from diverse analogical reasoning scenarios.

The finding that the UTD-NB system outperforms Llama on analogical reasoning tasks may potentially arise from its distinct utilization of unsupervised learning in identifying patterns between word pairs. By employing unsupervised learning, the LLM system can autonomously discover and discern intricate patterns within its subcategories without the need for labeled data. This approach allows the system to potentially capture more nuanced and detailed relationships between words, leading to a richer understanding of semantic associations. Consequently, this enhanced capability in pattern recognition and ranking could translate into improved performance on analogical reasoning tasks compared to Llama, which may rely on different methodologies or lack the same level of pattern recognition sophistication.

Beyond this presented analysis, comprehending the varied performance of Llama 2 across different sub-categories in a relational similarity task requires a deeper understanding of semantics from experts.

6 Conclusion

In conclusion, this study investigated the capabilities of Llama 2 in analogical reasoning through the framework of "SemEval-2012 Task 2: Measuring Degrees of Relational Similarity." Despite its limitations as a foundation model, Llama 2 demonstrated promising potential in discerning relational nuances. However, it fell short of outperforming the established models in certain subcategories, highlighting areas for further refinement and development. The incorporation of reversals in the dataset presented a unique challenge, testing the model’s depth of semantic comprehension. While Llama 2’s performance varied across different relational categories, it excelled in some areas, indicating its evolving proficiency in analogical reasoning tasks. Future research could focus on fine-tuning Llama 2 for specific analogical reasoning scenarios or expanding the scope by including the full testing set of subcategories. This study’s insights contribute to the broader understanding of AI’s capabilities in cognitive tasks and level of intelligence, specifically within the topic of analogical reasoning, paving the way for further advancements in AI research.

7 Appendix

7.1 A Taxonomy of Semantic Relations

In the context of our exploration into analogical reasoning, we present a comprehensive taxonomy of semantic relations. This taxonomy delineates ten distinct families or classes of relations, each comprising various specific relations as members. Below, we provide a concise characterization of each class along with their corresponding subcategories:

1. *Class Inclusion*: One word denotes a class that encompasses the entity named by the other word.

- a. Taxonomic - flower:tulip, emotion:rage
- b. Functional - ornament:brooch, weapon:knife
- c. Singular Collective - cutlery:spoon, clothing:shirt
- d. Plural Collective - groceries:eggs, dishes:saucers
- e. Class Individual - queen:Elizabeth, river:Nile

2. *Part-Whole*: One word refers to a part of the entity named by the other word, or something characteristic that is not a part.

- a. Object:Component: car:engine, face:nose
- b. Collection:Member: forest:tree, anthology:poem
- c. Mass:Portion: water:drop, mile:yard
- d. Event:Feature - rodeo:cowboy, banquet:food
- e. Activity:Stage - shopping:buying
- f. Item:Topological Part - room:corner, mountain:foot
- g. Object:Stuff - glacier:ice, parquet:wood
- h. Creature:Possession - millionaire:money, author:copyright
- i. Item:Distinctive Nonpart - tundra:tree, horse:wings,
- j. Item:Ex-part/Ex-possession - metal:dross, apostate:belief

3. *Similar*: One word represents a different degree or form of the object, action, or quality represented by the other word.

- a. Synonymity - car:auto, buy:purchase
- b. Dimensional Similarity - breeze:gale, enthusiasm:fervor

-
- c. Dimensional Excessive – eating:gluttony, walk:swagger
 - d. Dimensional Naughty – copy:plagiarize, listen:eavesdrop
 - e. Conversion – apprentice:master, colt:horse, grape:wine
 - f. Attribute Similarity – rake:fork, painting:movie
 - g. Coordinates – ram:ewe, son:daughter
 - h. Change – crescendo:sound
4. *Contrast*: One word denotes an opposite or incompatible aspect of the other word.
- a. Contradictory – alive:dead, fertile:sterile,
 - b. Contrary – old:young, happy:sad
 - c. Reverse – attack:defend, buy:sell
 - d. Directional – front:back, left:right
 - e. Incompatible – happy:morbid, frank:hypocritical
 - f. Asymmetric Contrary – hot:cool, dry:moist
 - g. Pseudoantonym – popular:shy, right:bad
 - h. Defective – default:payment, stutter:speech
5. *Attribute*: One word denotes a characteristic quality, property, or action of the entity named by the other word.
- a. Item:Attribute (noun:adjective) – beggar:poor, idyll:carefree
 - b. Object Attribute:Condition (adjective:adjective) – brittle:broken, malleable:molded
 - c. Object:State (noun:noun) – beggar:poverty, dupe:credulity
 - d. Agent Attribute:State (adjective:noun) – contentious:quarrels, taciturn:silence
 - e. Object:Typical Action (noun:verb) – glass:break, sycophant:flatter
 - f. Agent/Object Attribute:Typical Action (adjective:verb) – (agent attribute) viable:live, (object attribute) salient:notice, mandatory:comply (agent/object attribute) mutable:change, brittle:break
 - g. Action:Action Attribute – creep:slow
 - h. Action:Object Attribute – sterilize:infectious, capture:elusive
 - i. Action:Resultant Attribute (verb:noun/adjective) – stipple:dots,riddle:holes

6. *Nonattribute*: One word denotes a quality, property, or action that is characteristically not an attribute of the entity named by the other word.

- a. Item:Nonattribute (noun:adjective) – harmony:discordant, bulwark:flimsy
- b. Object Attribute:Noncondition (adjective:adjective) – brittle:molded, inconsolable:comforted
- c. Object:Nonstate (noun:noun) – laureate:honor, famine:plenitude
- d. Attribute:Nonstate (adjective:noun) – dull:cunning, immortal:death,
- e. Object:Atypical Action (noun:verb) – recluse:socialize, ascetic:indulge
- f. Agent/Object Attribute: Atypical Action (adjective:verb) – (agent attribute) reticent:talk, abstemious:gorge (object attribute) obtrusive:ignore, garble:comprehend
- g. Action:Action Nonattribute – creep:fast, fade:abruptly, scream:quietly, destroy:gently
- h. Action:Object Nonattribute – embellish:austere, obliterate:extant

7. *Case Relation*: One word denotes an action that the entity named by the other word is usually involved in, or both words name entities that are normally involved in the same action in different ways.

- a. Agent:Object – (product) tailor:suit, oracle:prophecy, jury:decision (raw material) baker:flour, sculptor:stone (associated object) plumber:pipe
- b. Agent:Recipient – doctor:patient, mentor:protege
- c. Agent:Instrument – farmer:tractor, conductor:baton
- d. Action:Object – plow:earth, baste:chicken
- e. Action:Recipient – bequeath:heir, teach:student
- f. Object:Recipient – inheritance:heir, speech:audience
- g. Object:Instrument – patient:stethoscope, water:sluice
- h. Recipient:Instrument – heir:testament

8. *Cause-Purpose*: One word represents the cause, purpose, or goal of the entity named by the other word, or the purpose or goal of using the entity named by the other word.

- a. Cause:Effect – enigma:puzzlement, joke:laughter
- b. Cause:Compensatory Action – hunger:eat, fatigue:sleep
- c. Enabling Agent:Object – match:candle, gasoline:car
- d. Action/Activity:Goal – eat:satiation, run:escape

- e. Agent:Goal – pilgrim:shrine, hunter:quarry
 - f. Instrument:Goal – anesthetic:numbness, ballast:stability
 - g. Instrument:Intended Action – gun:shoot, pestle:mash
 - h. Prevention – pesticide:vermin, splint:mobility
9. *Space-Time*: One word denotes a thing or action associated with a particular location or time named by the other word.
- a. Item:Location – arsenal:weapon, seminary:theologian
 - b. Location:Process/Product – bakery:bread, school:learning
 - c. Location:Action/Activity – school:learn, gym:exercise
 - d. Location:Instrument/Associated Item – school:textbook, farm:tractor
 - e. Contiguity – coast:ocean, sidewalk:road
 - f. Time:Action/Activity – summer:harvest, childhood:play
 - g. Time:Associated Item – retirement:pension, infancy:cradle
 - h. Sequence – prologue:narrative, inception:development
 - i. Attachment – hackles:neck, belt:waist, rivet:girder
10. *Reference*: One word denotes something that is an expression or representation of, or a plan or design for, or provides information about, the entity named by the other word.
- a. Sign:Significant – siren:danger, scepter:authority
 - b. Expression – smile:friendliness, lamentation:grief
 - c. Representation – person:portrait, backdrop:vista
 - d. Plan – agenda:meeting, procedure:flowchart
 - e. Knowledge – ornithology:birds, psychology:mind
 - f. Concealment – alias:name, code:meaning

8 References

- [1] AI Breakdown or: Takeaways From the 78-page Llama-2 Paper. (n.d.). Retrieved February 7, 2024, from <https://deepgram.com/learn/llama-2-paper-explained>
- [2] Bartha, P. (2022). Analogy and Analogical Reasoning. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2022/entries/reasoning-analogy/>
- [3] Jurgens, D., Mohammad, S., Turney, P., & Holyoak, K. (2012). SemEval-2012 Task 2: Measuring Degrees of Relational Similarity. In E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Eds.), **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)* (pp. 356–364). Association for Computational Linguistics. <https://aclanthology.org/S12-1047>
- [4] Korteling, J. E. (Hans)., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence*, 4. <https://www.frontiersin.org/articles/10.3389/frai.2021.622364>
- [5] Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10), 4176–4181. <https://doi.org/10.1073/pnas.1814779116>
- [6] Ph.D, C. R. W. (2023, July 11). LLaMA: LLMs for Everyone! Medium. <https://towardsdatascience.com/llama-llms-for-everyone-724e737835be>
- [7] Rehan, A. (2023, September 6). Llama 2 Explained in Detail within 5 Minutes. Geekflare. <https://geekflare.com/llama-2-explained/>
- [8] Timothy, M. (2023, July 20). What Is Llama 2 and How Can You Use It? MUO. <https://www.makeuseof.com/what-is-llama-2-and-how-can-you-use-it/>
- [9] What is Llama 2 and why does it matter? (n.d.). Retrieved February 7, 2024, from <https://zapier.com/blog/llama-meta/>