Università degli Studi di Padova

Analogical Reasoning with Llama 2

Course: Cognitive, Behavior, and Social Data 2023-2024
Group 12: Nour Al Housseini, Sofia Trogu, Fairouz Baz Radwan

Presentation Key Words

➔ *Relational class:* Defines relation between two words.

➔ *Instances:* Word pairs included in a relation class **.

➔ *Turkers:* Humans working as Amazon Mechanical Turk Worker.

➔ *MaxDiff:* Questions used in Phase 2 of the study to determine most and least illustrative among a set of word pairs.

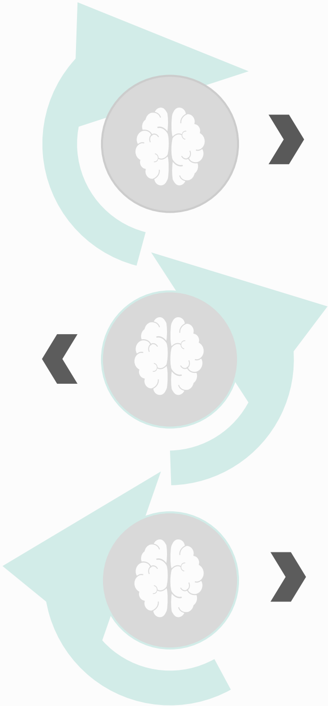
➔ *Prototypicality:* Percentage of times a word pair is chosen as most illustrative minus the times it's chosen as least illustrative based on MaxDiff answers.

➔ *Spearman Correlation:* Rank correlation coefficient metric used to measure strength of relationship between an AI system's and human's prototypicality ratings.

Paper Summary

- ➔ Goal of study: measure relational similarity among instances in a relation class.
  - ◆ Instances have high relational similarity if they closely express the same relationship in a relation class.
- ➔ Study was performed in 2 Phases:
  - ◆ Phase 1: Generation of word pairs in each relation class by Turkers.
  - ◆ Phase 2: Ranking of pairs from Phase 1 to produce MaxDiff prototypicality scores.
- ➔ Other parts to the study and metrics:
  - ◆ Comparison of Turkers' results to 6 different AI systems
  - ◆ Spearman correlation
  - ◆ Inclusion of reversals for noise injection
- ➔ Findings: Certain systems achieved moderate performance in some subcategories, but no system was able to achieve superior performance overall.

Systems Used for Comparison



UMD-V2

Same procedure as V0, with two further expansions to related concepts.

UMD-V0

WordNet is used to build the set of concepts connected by WordNet relations to the pairs' words. Prototypicality is estimated using the vector similarity of the concatenated glosses.

UTD-NB

Unsupervised learning identifies intervening patterns between all word pairs, consequently ranked according to subcategory specificity.

UMD-V1

Same procedure as V0, with one further expansion to related concepts.

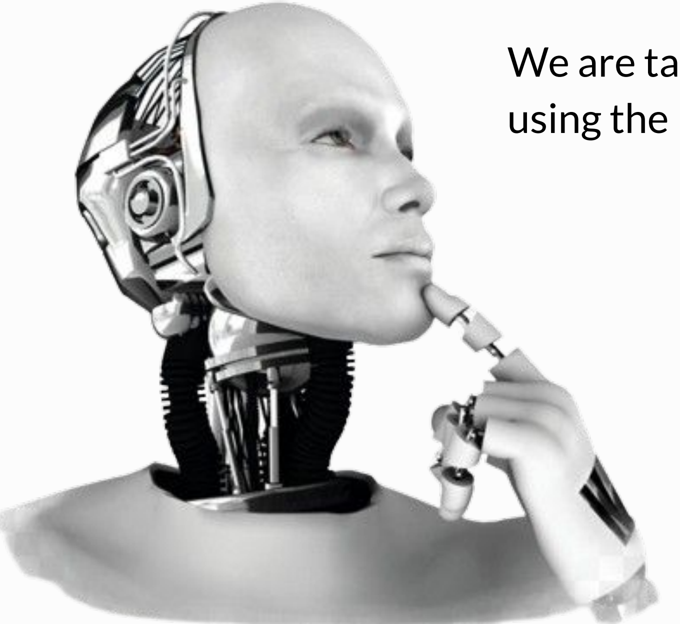
UTD-SVM

Intervening patterns are found using the same method as UTD-NB. Word pairs are then represented as feature vectors of matching patterns.

BUAP

Each pair is represented as a vector over multiple features: lexical, intervening words, WordNet relations, and syntactic features. Prototypicality is based on cosine similarity with the class's pairs.

Our Objective



We are tasked with re-creating Phase 2 of the study given to the Turkers, using the large language model Llama 2 and measuring its performance.

Llama 2 Task

Example:

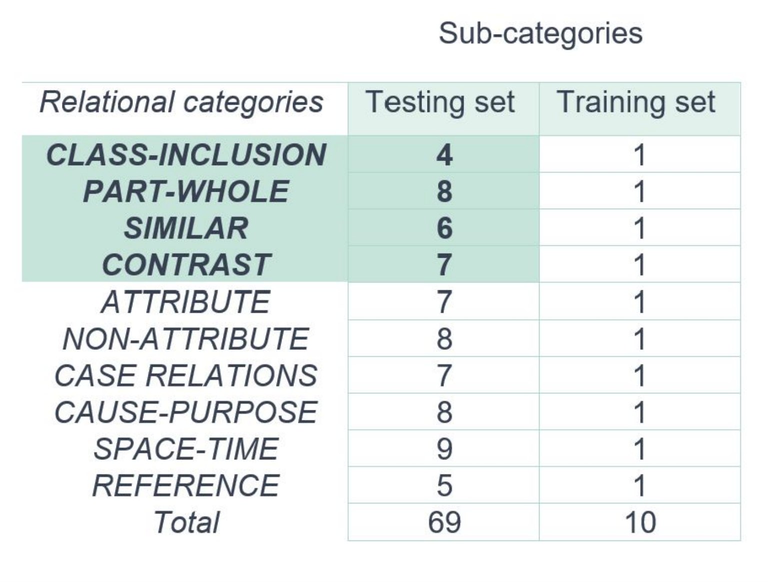
Category: Class-Inclusion: Functional
Relation: Y functions as an X
Word Pairs:
“fridge:appliance”,”chisel:tool”,”preservative:salt”,”seat:chair”

To conduct our study, we instructed our LLM to distinguish among pairs of relation instances generated by Turker’s in Phase 1; based on how representative the examples are to a relation category.

For each unique combination of 4 instances (of each subcategory), we provided a general prompt that included the relation followed by the four specific pairs and asked the model to give the most and least illustrative pair.

Datasets & Subsetting

- ➔ The dataset utilized in our study was constructed using surveys carried out by the paper's authors on human participants.
- ➔ The dataset was composed of ten main relational categories further divided into 79 subcategories, with ten allocated for training and 69 for testing.
- ➔ Our analysis concentrated on four main categories highlighted in the table, encompassing 25 subcategories in the testing set.
- ➔ For each subcategory, we dealt with approximately 100 unique sets of four-word pairs.

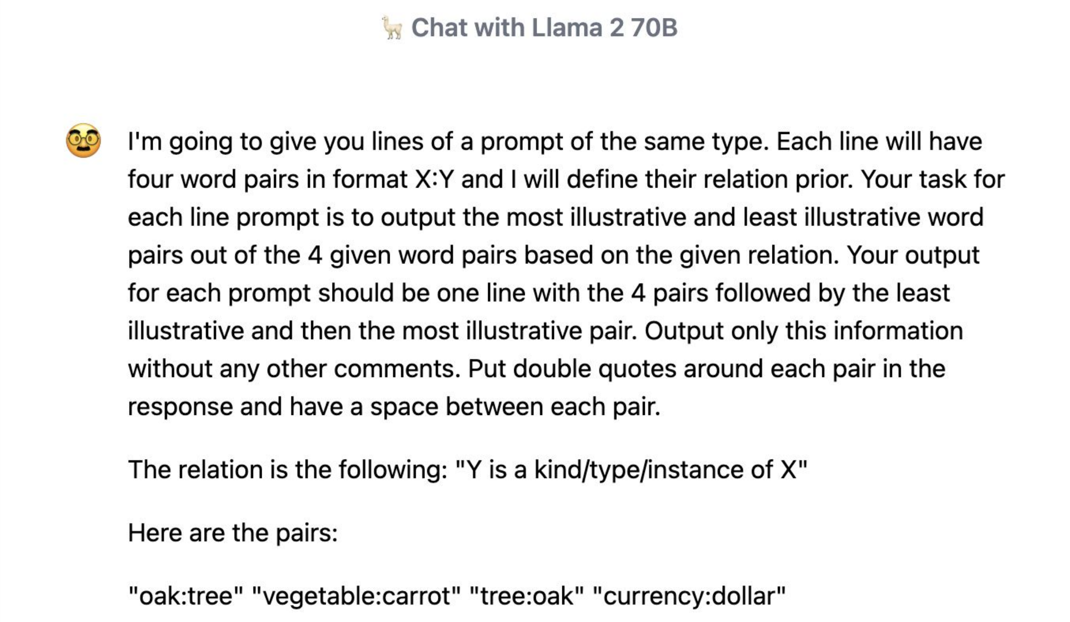


	Sub-categories	
Relational categories	Testing set	Training set
CLASS-INCLUSION	**4**	1
PART-WHOLE	**8**	1
SIMILAR	**6**	1
CONTRAST	**7**	1
ATTRIBUTE	7	1
NON-ATTRIBUTE	8	1
CASE RELATIONS	7	1
CAUSE-PURPOSE	8	1
SPACE-TIME	9	1
REFERENCE	5	1
Total	69	10

Prompt formation

- ➔ 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”.
- ➔ 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.
- ➔ We tested inputting multiple sets of pairs in each prompt and observed adverse effects on the model's responses, and so we transitioned to a single-set input method. A more time-consuming approach but one that significantly improved results.

Our Final Prompt for Llama 2



🦙 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.

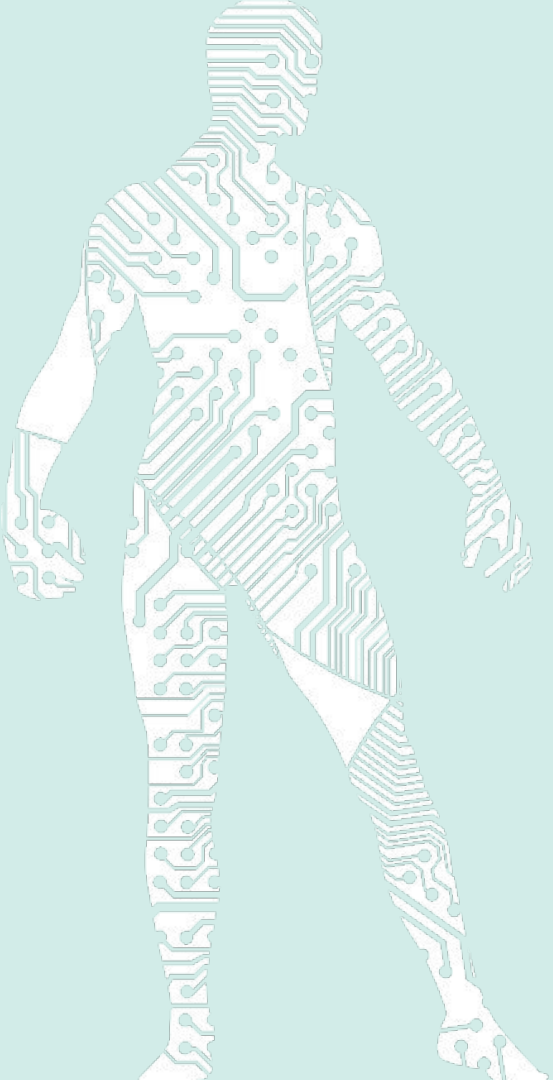
The relation is the following: "Y is a kind/type/instance of X"

Here are the pairs:

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

Adding 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.
- ➔ By embedding these reversals, we can analyse the LLMs' depth of understanding and their ability to navigate complex relational dynamics.
- ➔ An example of such word pairs in subsection 1a: tree:oak --> oak:tree.



Preliminary 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.
- ➔ We found out:
  - ◆ 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).
  - ◆ 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.

Performance Metrics

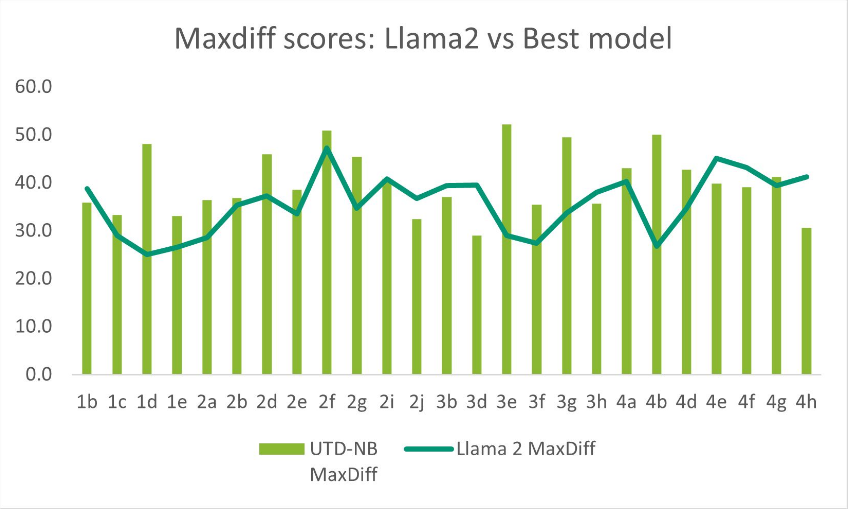
➔ MaxDiff:

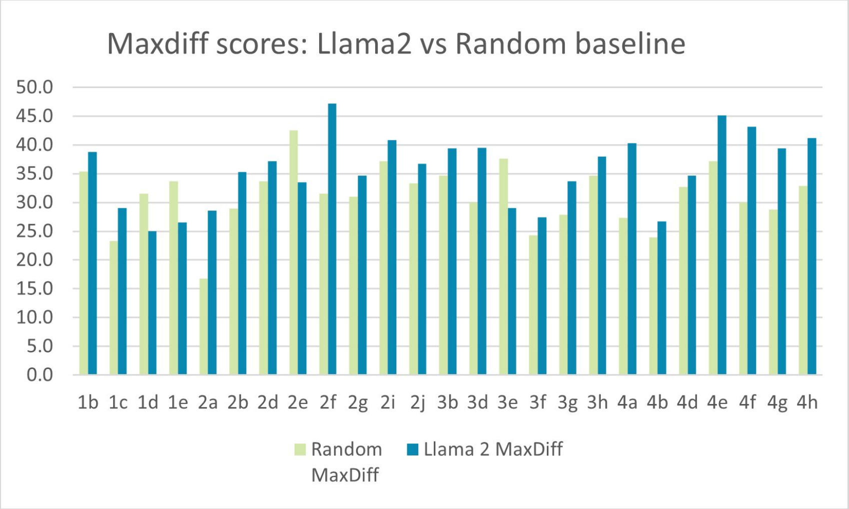
A key metric indicating the accuracy of the model’s responses for the MaxDiff questions calculated for each subcategory . This score served as a primary indicator of performance.

➔ Spearman Correlation:

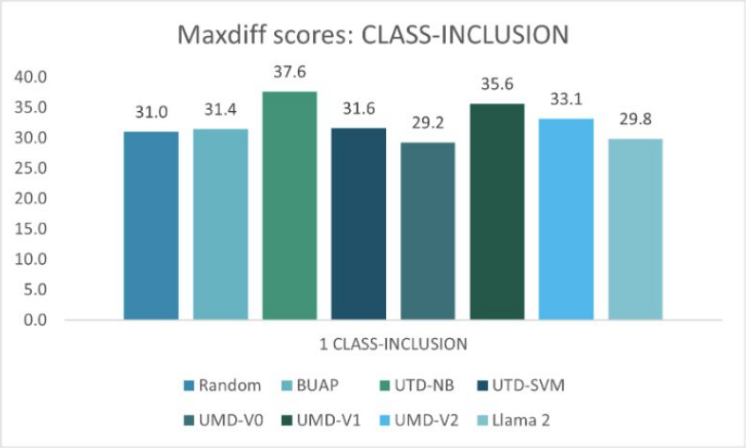
Correlation calculated between our model's prototypicality rating of word pairs and the ratings produced by the Turkers used to measure the degree of their alignment.

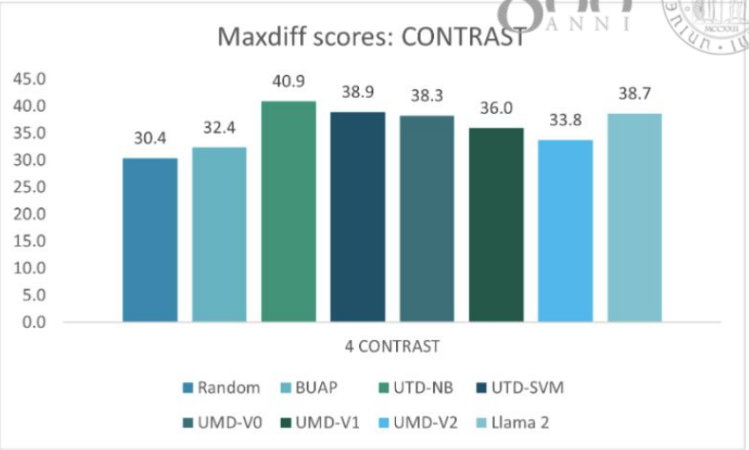
Analytical Differences:

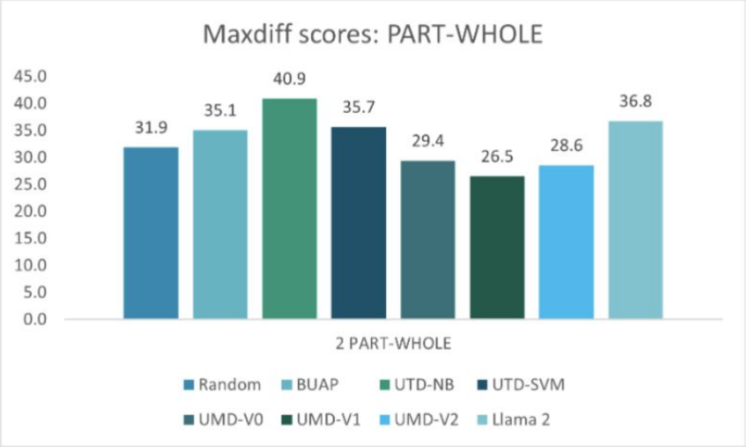


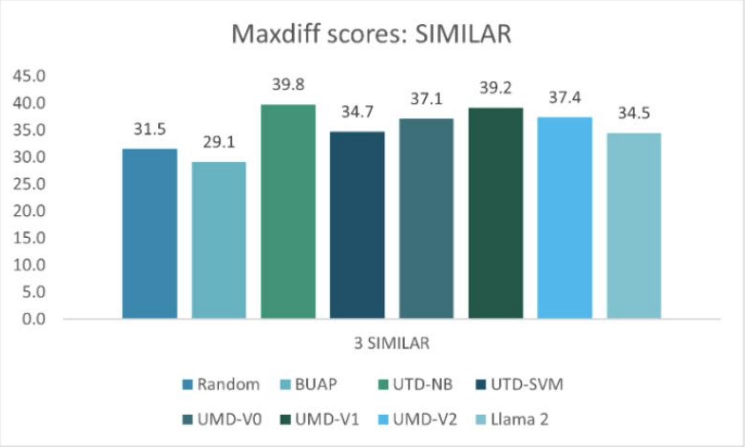


Average MaxDiff for the 4 subcategories





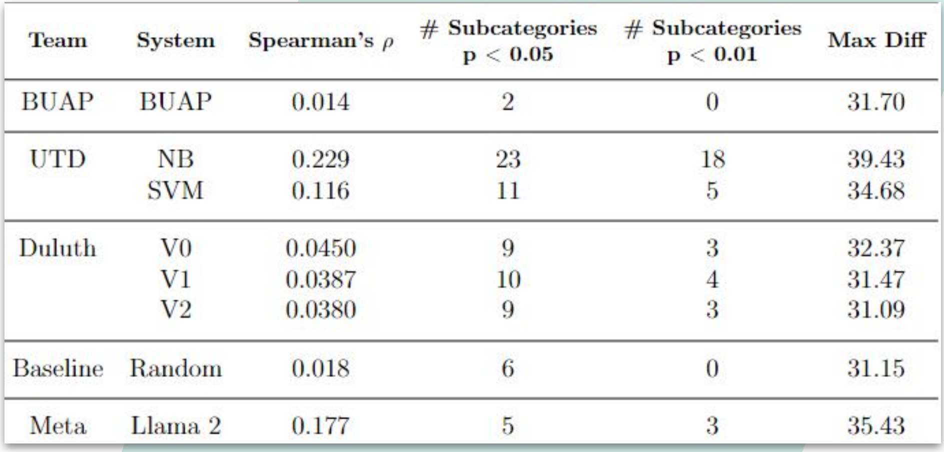




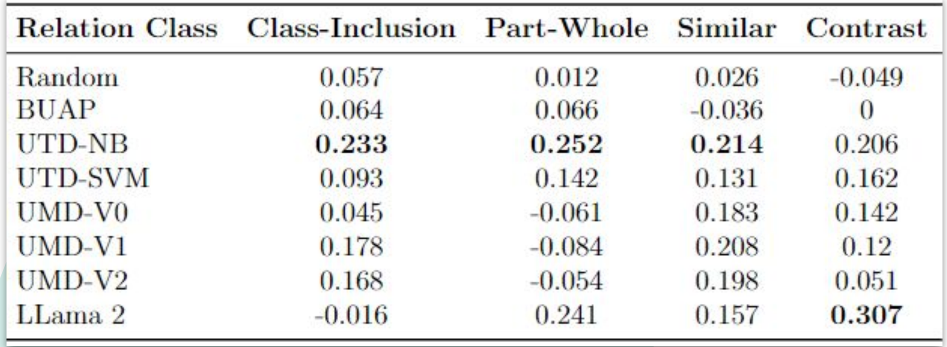
An important perspective to consider is the comparisons of the different systems.

The first table shows the average Spearman and MaxDiff scores for all the systems across the 69 test subcategories. Column 4 and 5 denote the number of subcategories with a statistically significant spearman.

The second table shows the average Spearman calculated with the Turker ratings in each of the relation categories.



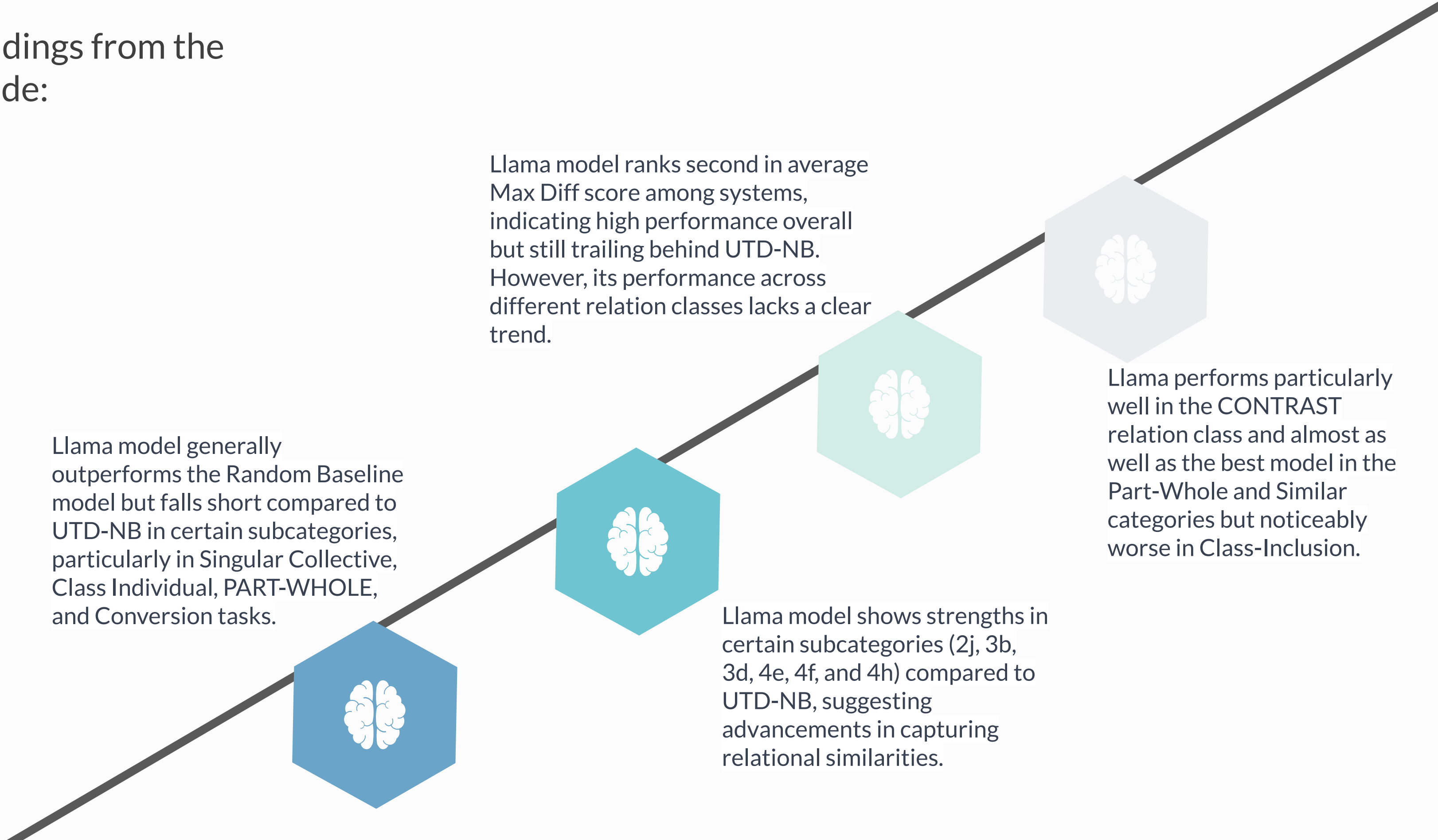
Team	System	Spearman's ρ	# Subcategories $p < 0.05$	# Subcategories $p < 0.01$	Max Diff
BUAP	BUAP	0.014	2	0	31.70
UTD	NB	0.229	23	18	39.43
	SVM	0.116	11	5	34.68
Duluth	V0	0.0450	9	3	32.37
	V1	0.0387	10	4	31.47
	V2	0.0380	9	3	31.09
Baseline	Random	0.018	6	0	31.15
Meta	Llama 2	0.177	5	3	35.43



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**

Discussion

Important findings from the analysis include:



Llama model generally outperforms the Random Baseline model but falls short compared to UTD-NB in certain subcategories, particularly in Singular Collective, Class Individual, PART-WHOLE, and Conversion tasks.

Llama model shows strengths in certain subcategories (2j, 3b, 3d, 4e, 4f, and 4h) compared to UTD-NB, suggesting advancements in capturing relational similarities.

Llama model ranks second in average Max Diff score among systems, indicating high performance overall but still trailing behind UTD-NB. However, its performance across different relation classes lacks a clear trend.

Llama performs particularly well in the CONTRAST relation class and almost as well as the best model in the Part-Whole and Similar categories but noticeably worse in Class-Inclusion.

Future Steps and Considerations

- ➔ Expand testing dataset to the rest of the subcategories
- ➔ Perform further investigation into effect of reversals
- ➔ With expansion of dataset and computational resources, exploy fine-tuning techniques

Thank You!

Any Questions?