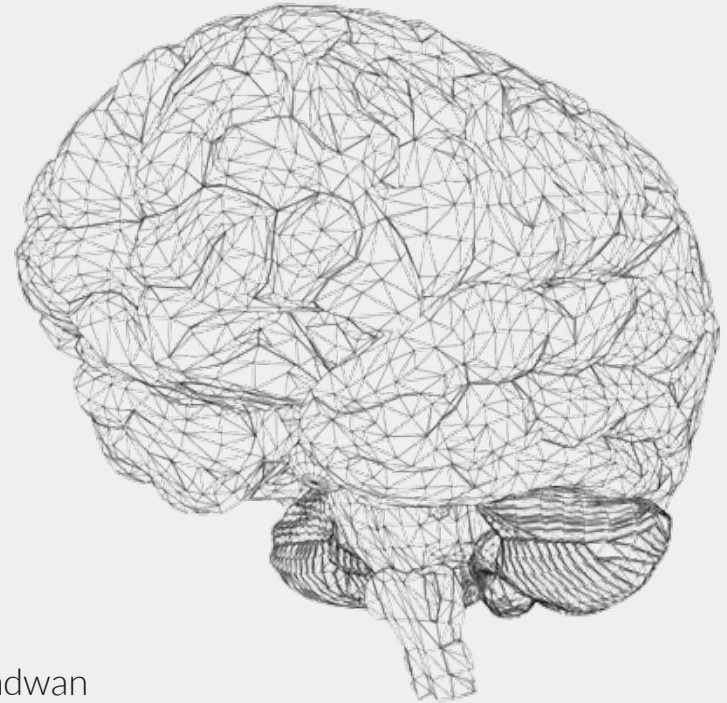
Analogical Reasoning with Llama 2



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

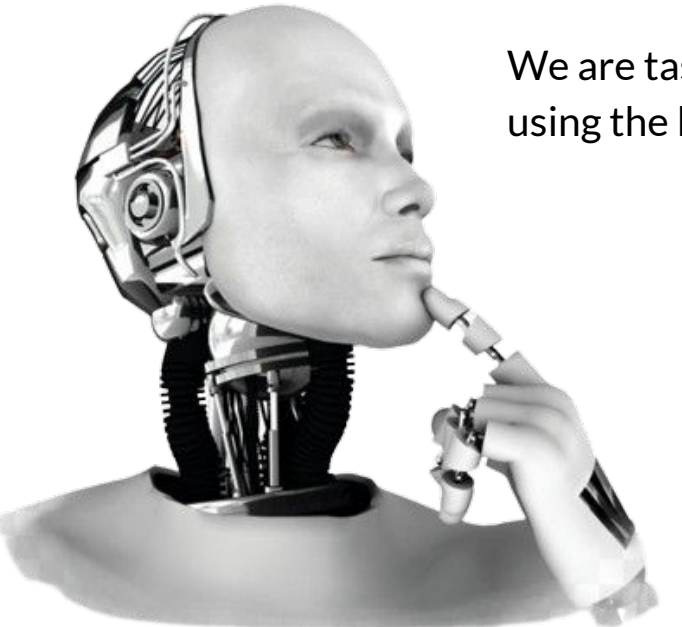
Presentation Key Words

- ➔ *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.

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.

Prompt Formation

🦙 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"

Datasets

Categories	Number of subcategories in the Testing Set
1- CLASS-INCLUSION	4
2- PART-WHOLE	8
3- SIMILAR	6
4- CONTRAST	7
5- ATTRIBUTE	7
6- NON-ATTRIBUTE	8
7- CASE RELATIONS	7
8- CAUSE-PURPOSE	8
9- SPACE-TIME	9
10- REFERENCE	5
Total	69

***Note*:**
Each subcategory has ~100 prompts, adding up to 6900 prompts that we have to provide to Llama 2, which we estimate will take about 23 hrs per person.

Preliminary Results

We completed the prompts for **6 subcategories** in testing. This adds up to 18 in total covering the first **3 categories** for which we computed the maxdiff accuracy scores and spearman correlation.

The following table shows the results we got for our model for 1 subcategory comparing it to the models presented in the paper.

Category 2: CLASS-WHOLE

Model	System	Maxdiff score	Spearman correlation
Llama 2	**Llama 2**	**36.8**	**0.241**
BUAP	BUAP	35.1	0.066
UTD	NB	**40.9**	**0.252**
	SVM	35.7	0.142
UMD	V0	29.4	-0.061
	V1	26.5	-0.084
	V2	28.6	-0.054

Future Steps and Considerations

- ➔ Additional data collection, if necessary.
- ➔ Analyze results for MaxDiff Accuracy scores and Spearman correlations.
- ➔ Determine possible reasons for model performance and understand limitations of Llama in context of human intelligence.

Thank You!

Any Questions?