Evaluating LLM performance on question-answering based on Wikidata triples

Aksel Mads Madsen

Introduction

- Questions answering
- Dataset contamination
- Answer rating difficulties
  - Quantify answer quality

Wikidata & question generation

- Large graph database
- Diverse Subjects
- Example of triple: `(Public Parks (Ireland) Act 1869 | legislated by | Parliament of the United Kingdom)`
- Filtering
- Question generating: 10-shot claude

Evaluation & Embeddings

- Language Models
  - GPT4-o
  - Claude 3 Sonnet
  - Mistral Large
- Embeddings
  - Semantic Similarity
  - Distances
- Bias potential
  - Multiple Embedding suppliers
- Rank Normalization

Analysis & Results

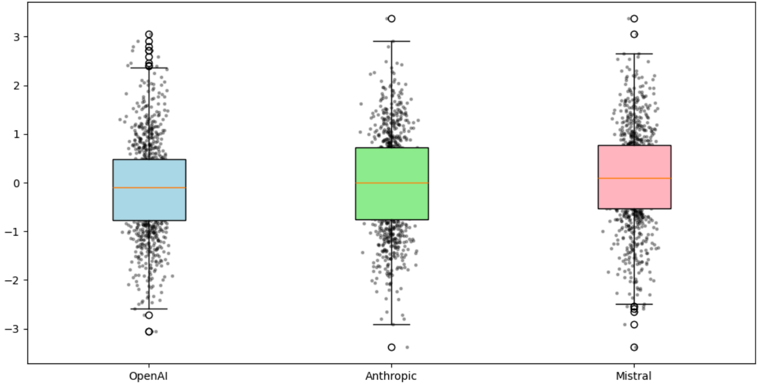
- Means of normalized distance

GPT	-0.101
Claude	0.004
Mistral	0.105

- ANOVA Results

Term	P-value	Effect size (F^2)
Model	$6.6 \cdot 10^{-5}$	$7 \cdot 10^{-3}$
Model-Rater Interaction	0.44	$5 \cdot 10^{-4}$

Results box-plot (In standard deviations)



Pos-hoc analysis

- Pos-hoc matrix
  - Tukey Post-Hoc test (in p-values):

	GPT4-o	Claude 3 Sonnet	Mistral Large
GPT4-o	1	0.095	0.001
Claude 3 Sonnet		1	0.05
Mistral Large			1

Limitations

- Small Effect size
- Poor dataset
- Embeddings distance difficult to interpret
  - Distance between embeddings between topic vary