LLM Evaluation

17 pages summary

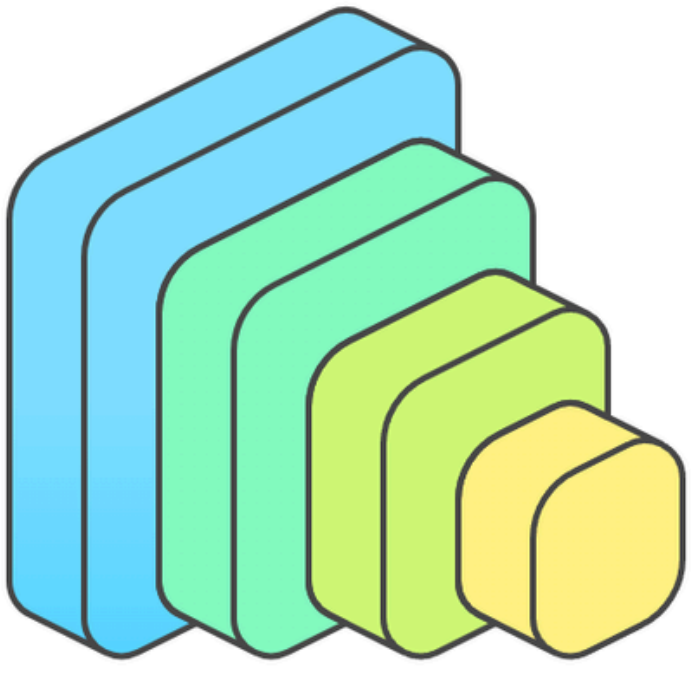
Refining Language Model Evaluation

Scenario and Metric Taxonomy

Organizing use cases and metrics

Multi-Metric Measurement

Evaluating models across multiple metrics



Targeted Evaluation

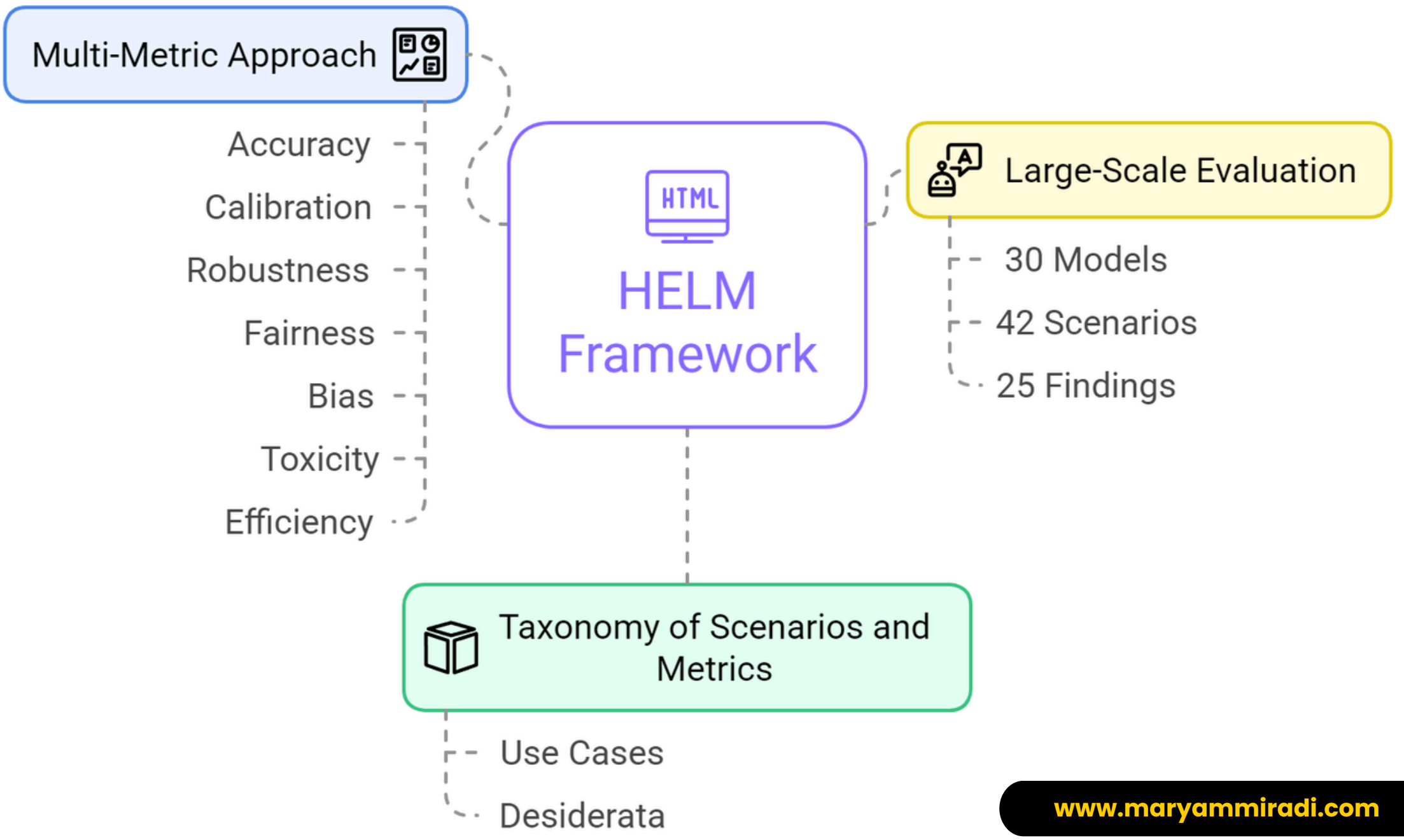
Deep analysis of specific aspects

Large-Scale Model Evaluation

Comprehensive assessment of models

LLM Evaluation

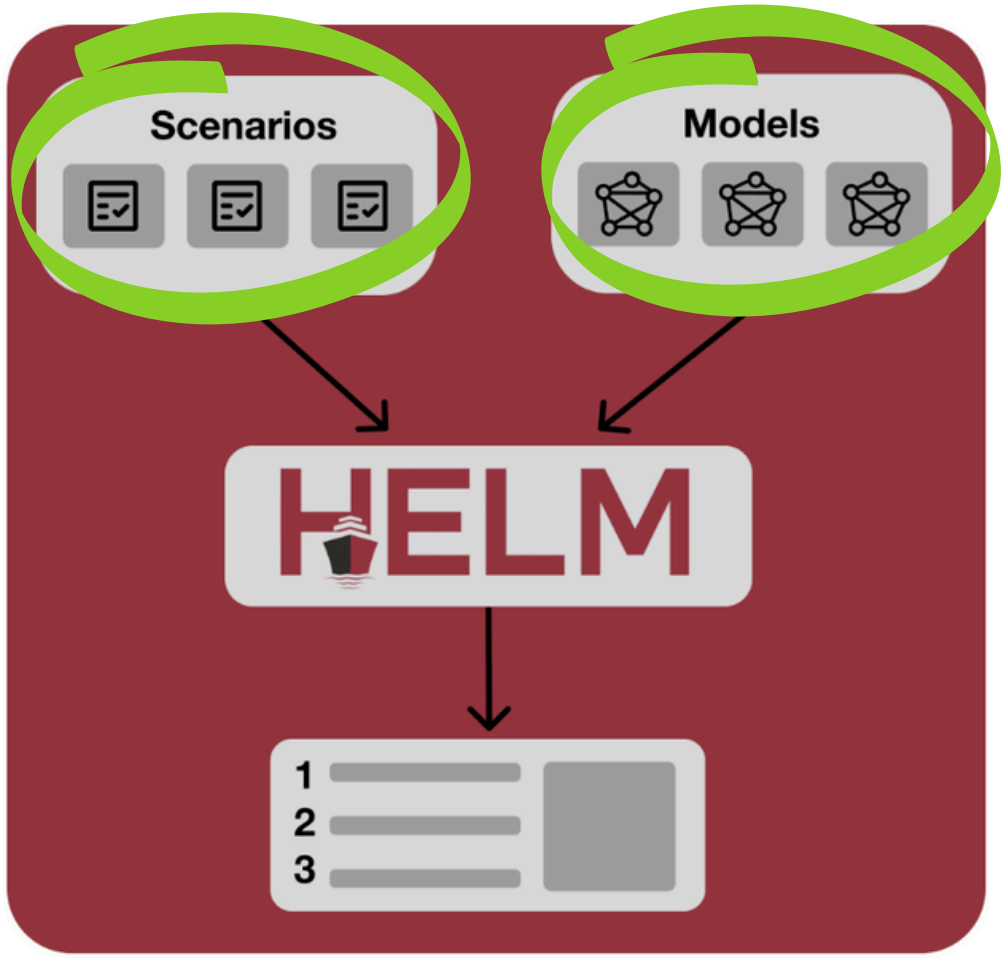
Overview HELM



LLM Evaluation

2/17

Winning Models HELM

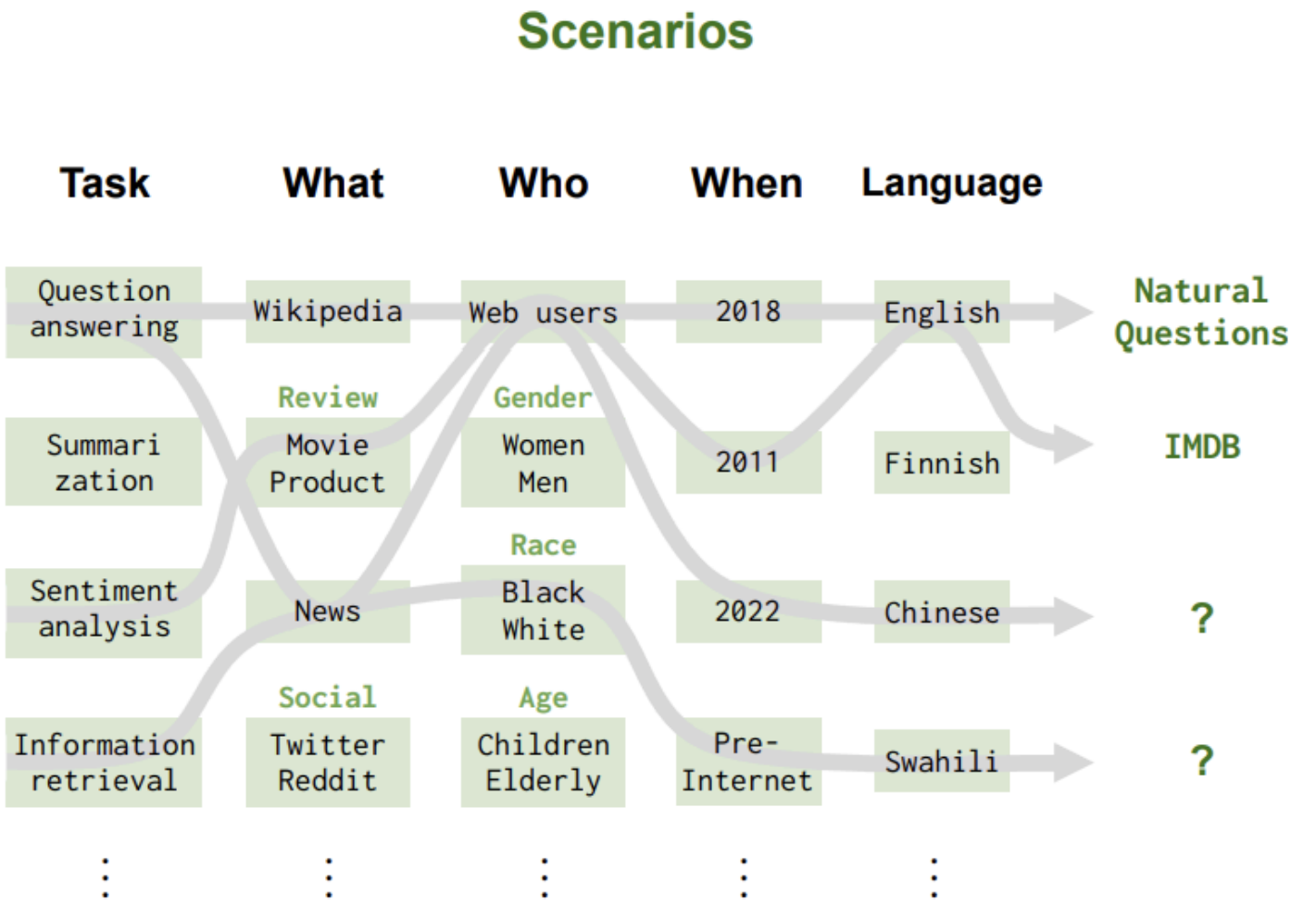


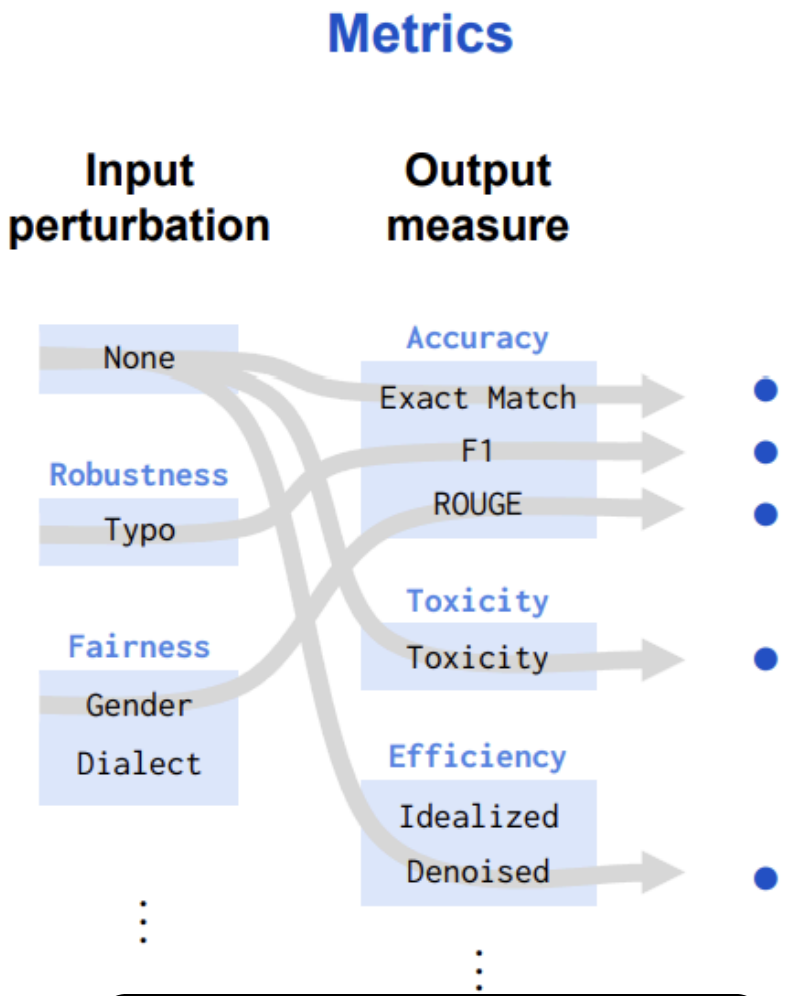
Model	Mean win rate
GPT-4o (2024-05-13)	**0.938**
GPT-4o (2024-08-06)	0.928
DeepSeek v3	0.908
Claude 3.5 Sonnet (20240620)	0.885
Amazon Nova Pro	0.885
GPT-4 (0613)	0.867
GPT-4 Turbo (2024-04-09)	0.864
Llama 3.1 Instruct Turbo (405B)	0.854
Claude 3.5 Sonnet (20241022)	0.846
Gemini 1.5 Pro (002)	

LLM Evaluation

3/17

Scenarios & Models





LLM Evaluation

Metrics

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✔	✔	✔	✔	✔	✔	✔
IMDB	✔	✔	✔	✔	✔	✔	✔
Natural Questions	✔	✔	✔	✔	✔	✔	✔
QuAC	✔	✔	✔	✔	✔	✔	✔
XSUM	✔				✔	✔	✔

LLM Evaluation

5/17

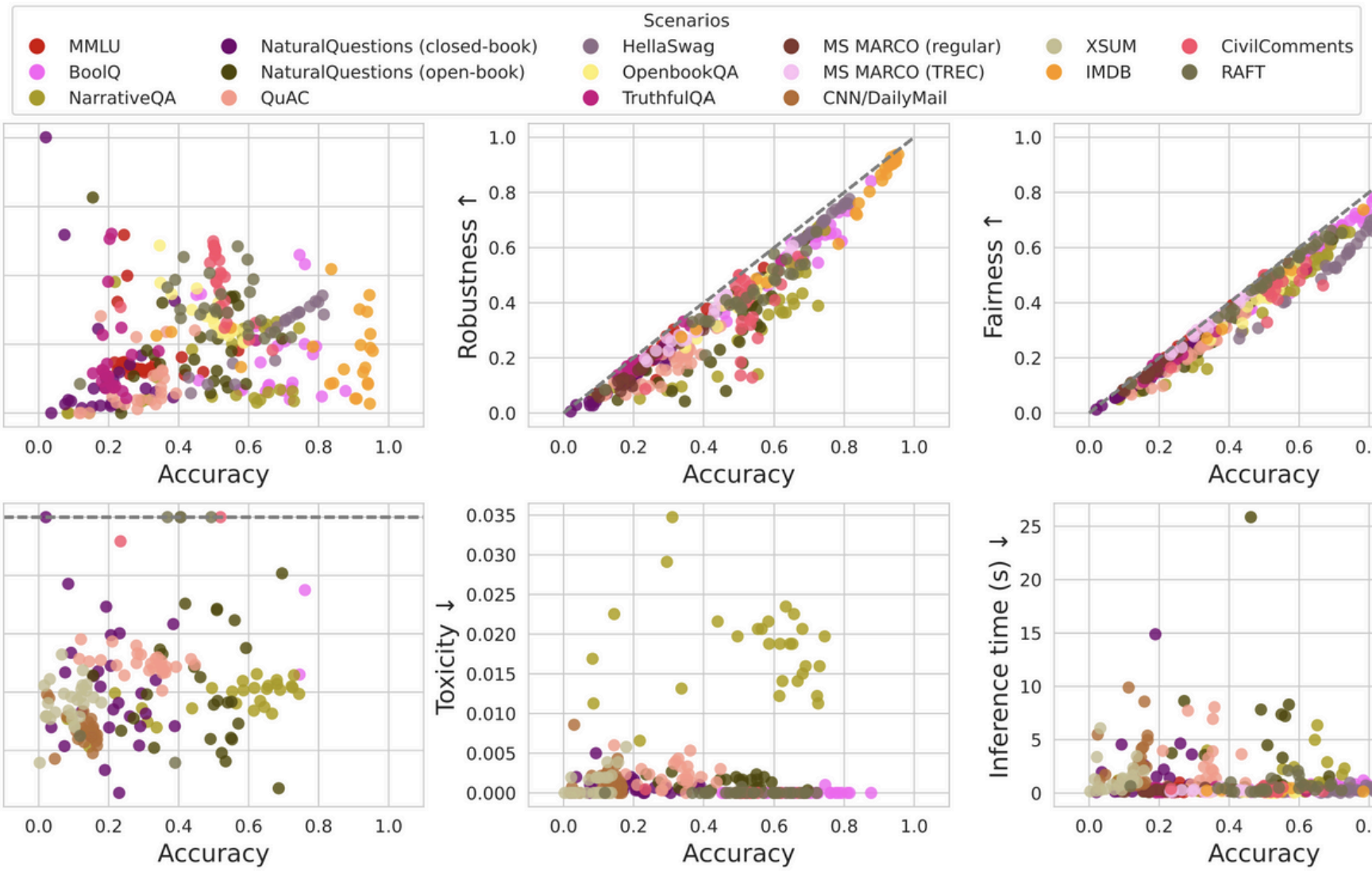
6000 World Languages

www.maryammiradi.com

LLM Evaluation

6/17

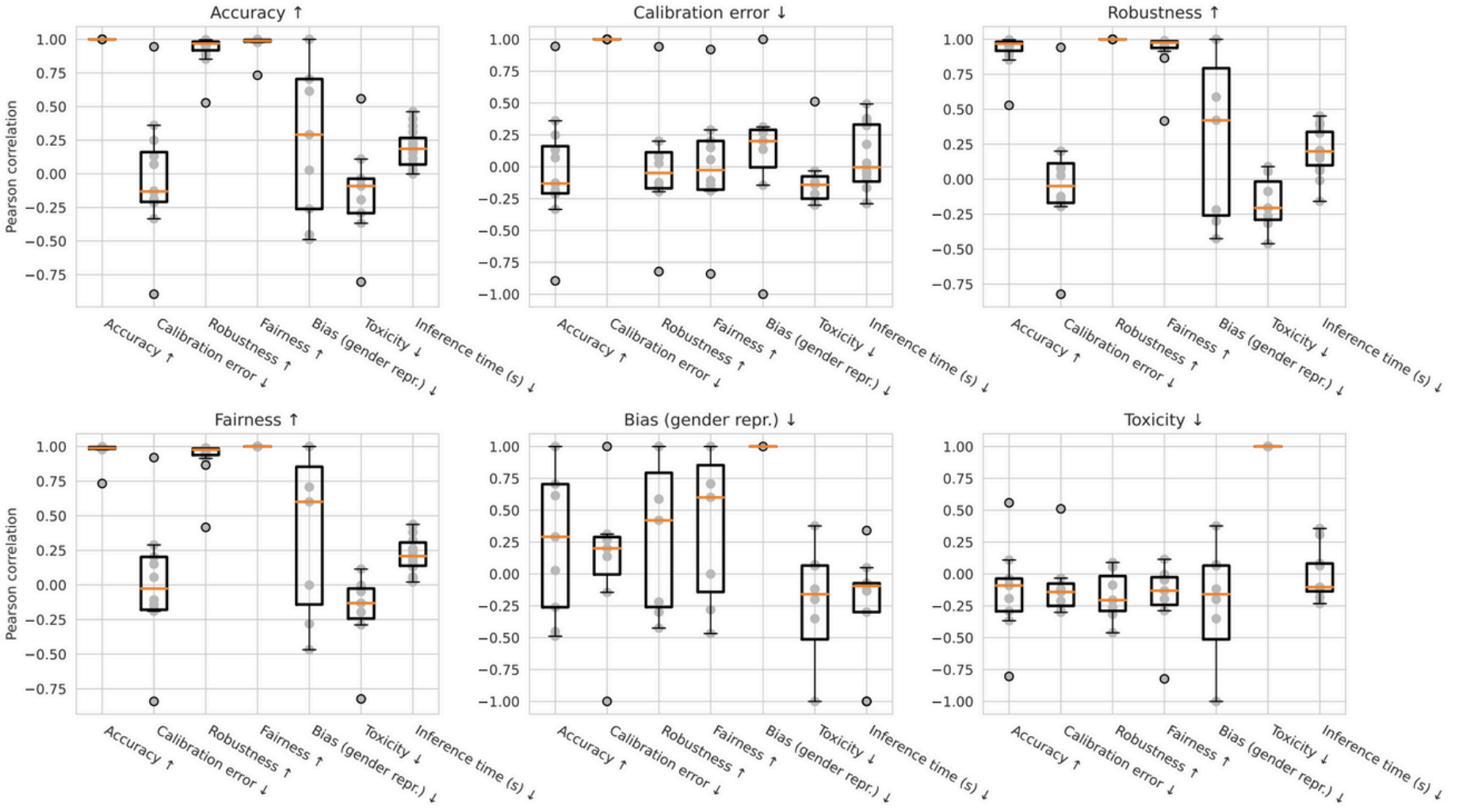
Accuracy of Metrics



LLM Evaluation

7/17

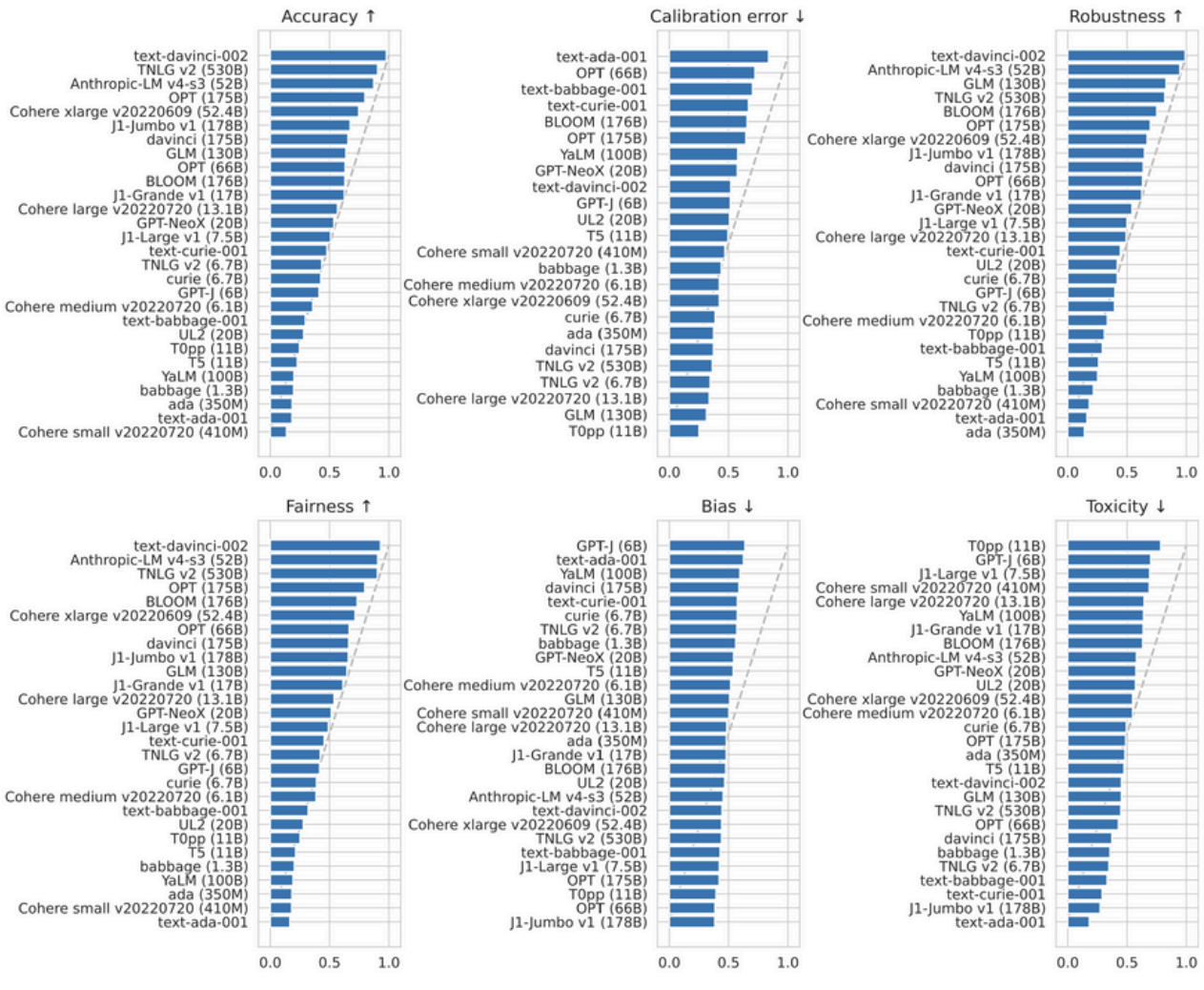
Correlation between Metrics



LLM Evaluation

8/17

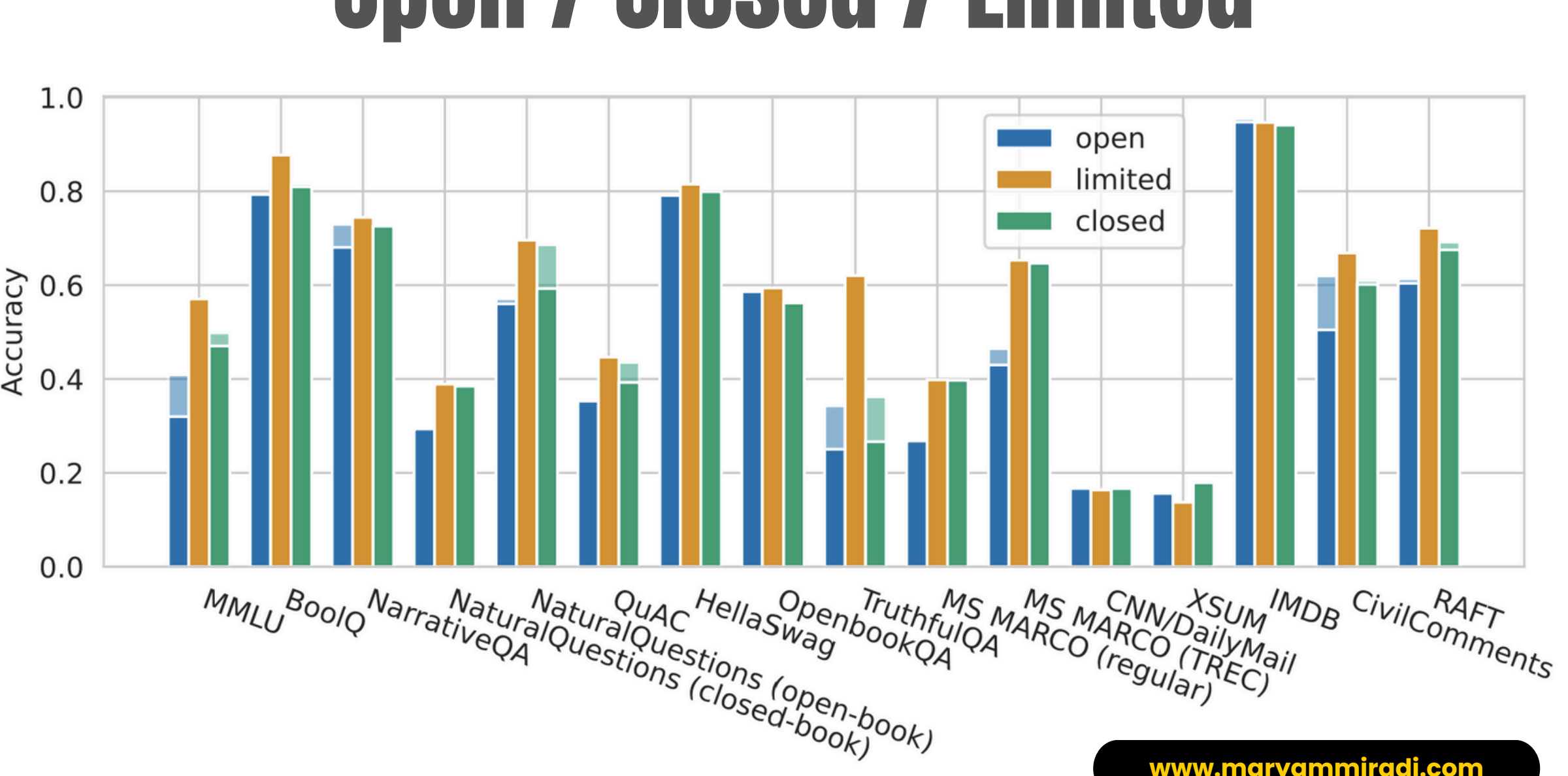
Win Rate per LLM



LLM Evaluation

9/17

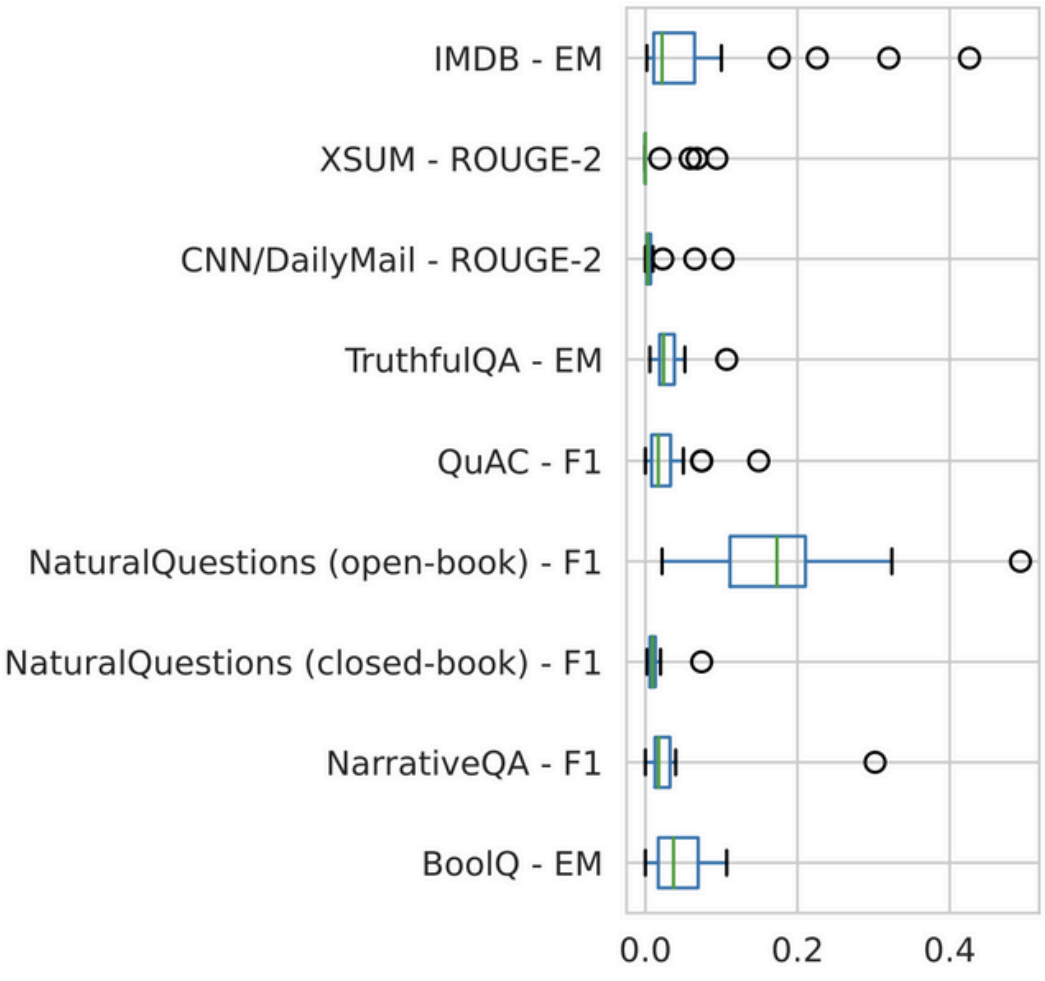
Open / Closed / Limited

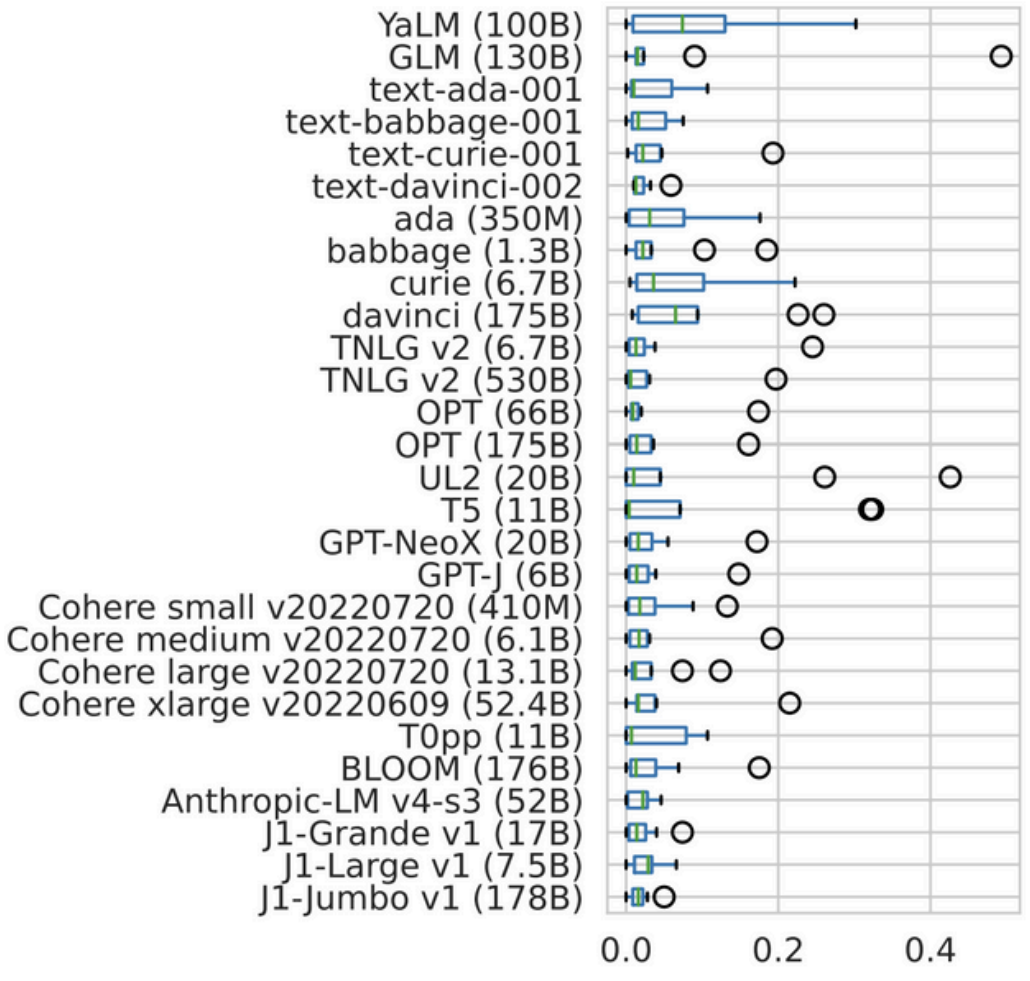


LLM Evaluation

10/17

Variance of Accuracy

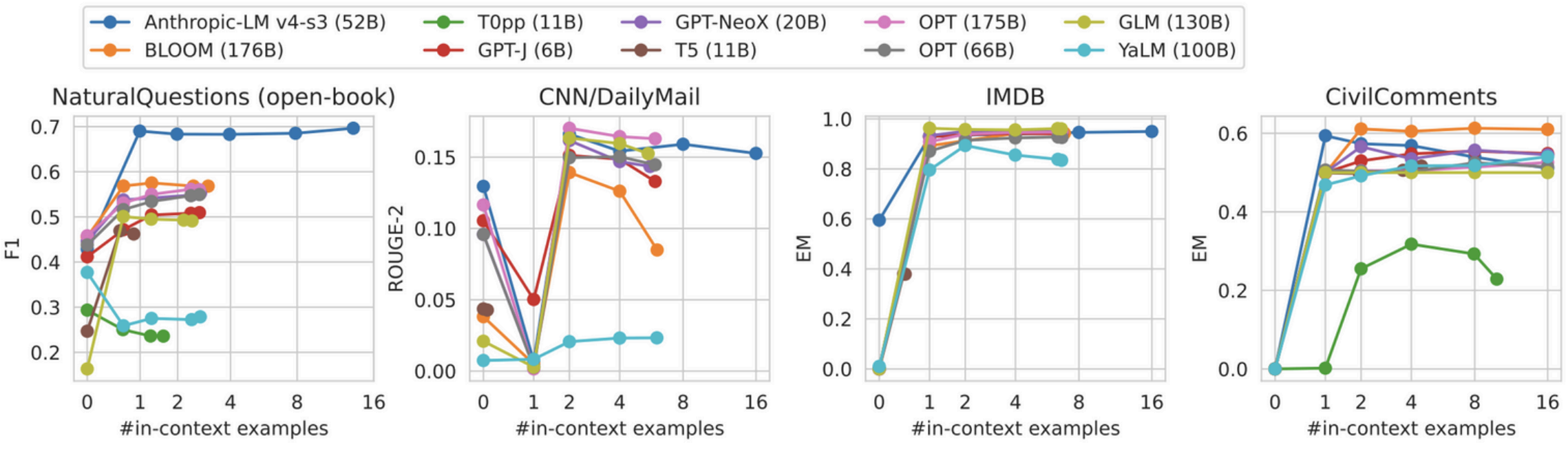




LLM Evaluation

11/17

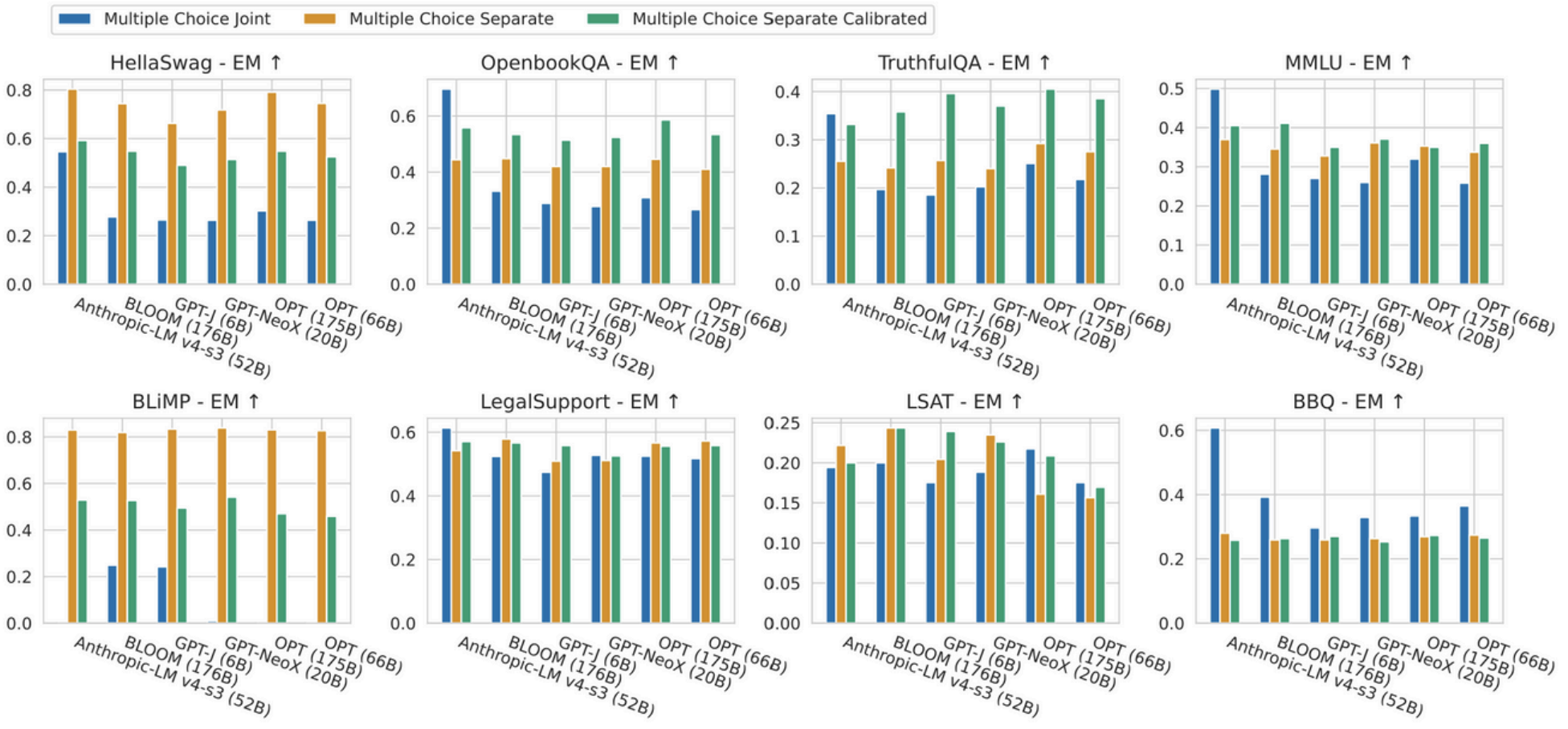
Number of in-Context Examples



LLM Evaluation

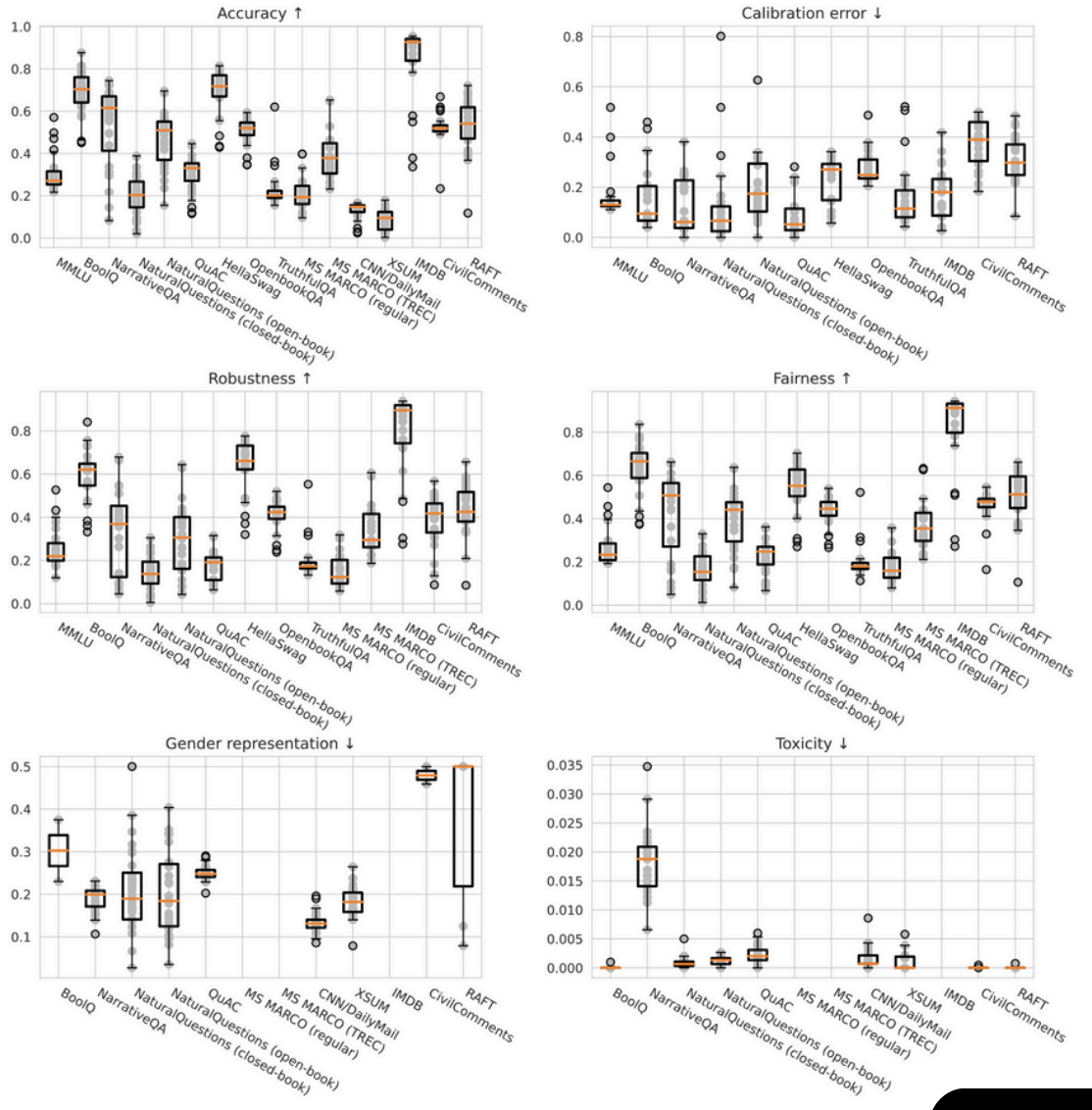
12/17

Multiple-choice adaptation



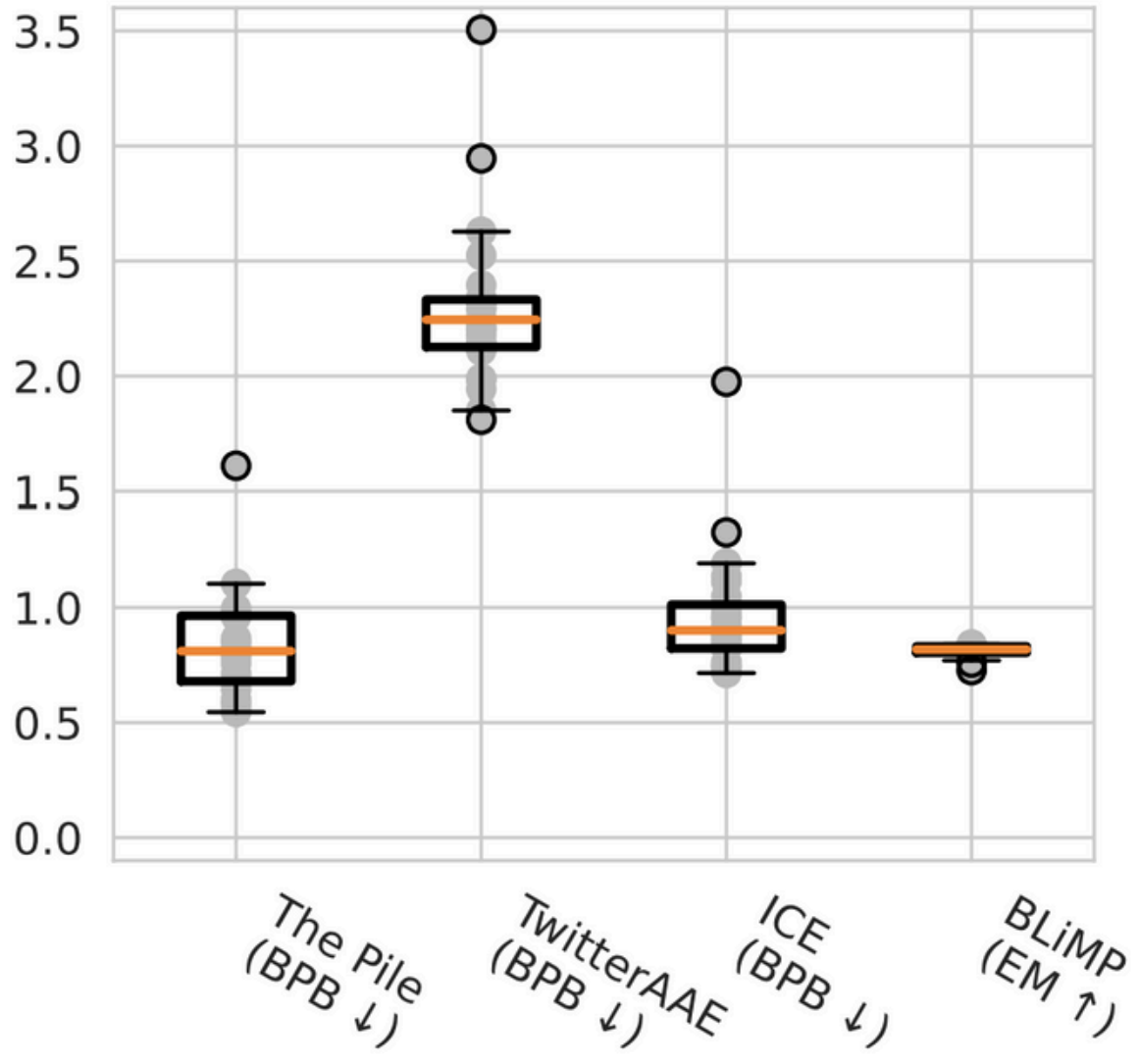
LLM Evaluation

Metric spread



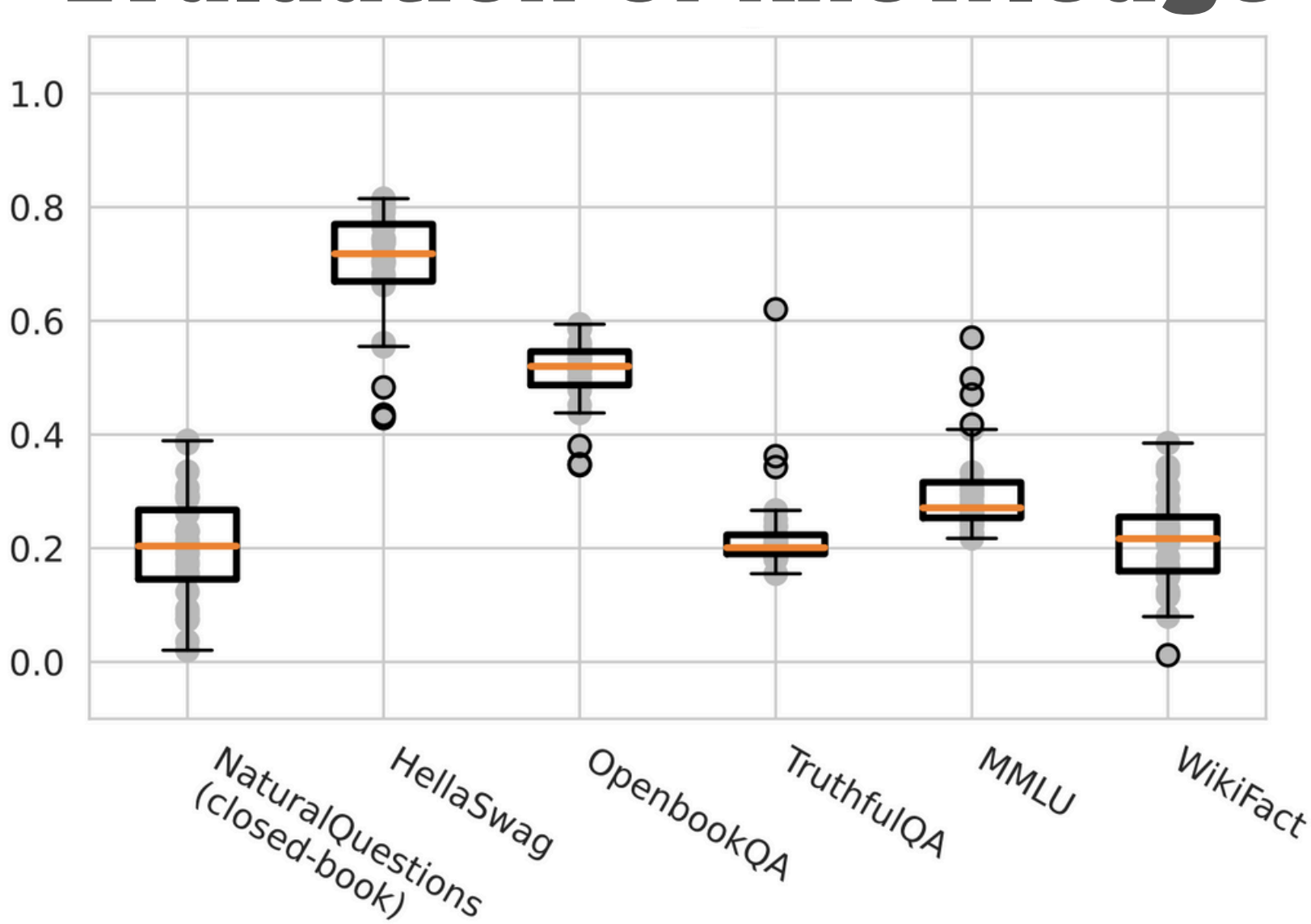
LLM Evaluation

Evaluation of Language



LLM Evaluation

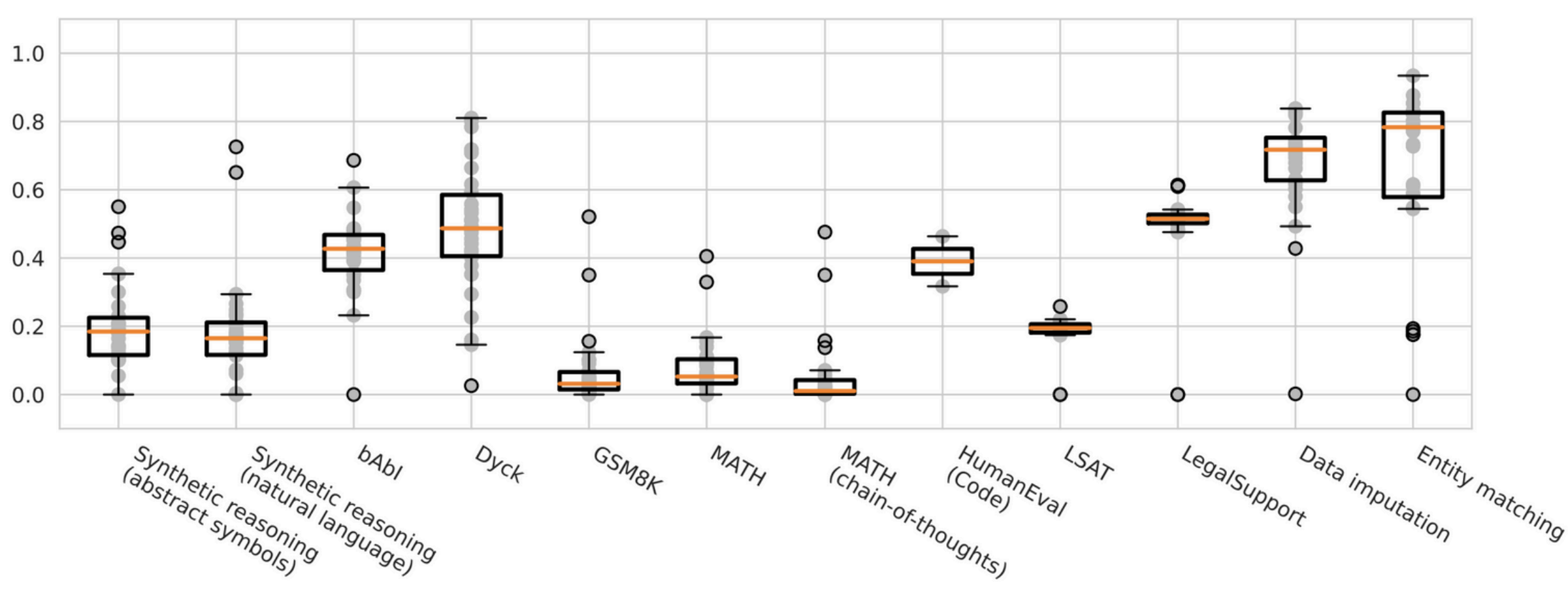
Evaluation of Knowledge



LLM Evaluation

16/17

Evaluation of Reasoning



LLM Evaluation

17/17

Evaluation of Social Bias