Evaluating Open Large Language Model Benchmark Platforms

Shane Ganz, Darasimi Ogunbinu-Peters,
Mentors: Mohammed Fakhruddin Babar , Monowar Hasan
EECS, Washington State University, Pullman, USA

INTRODUCTION

- The HELM (Holistic Evaluation of Language Models) test was developed by the Stanford University's Center for Research on Foundational Models as a benchmark to transparently evaluate language models.

Figure 1: HELM Logo

- It defines a framework to organize all possible metrics of evaluation. This allows HELM to measure the efficacy of what the model can do, and also identify what it can't.

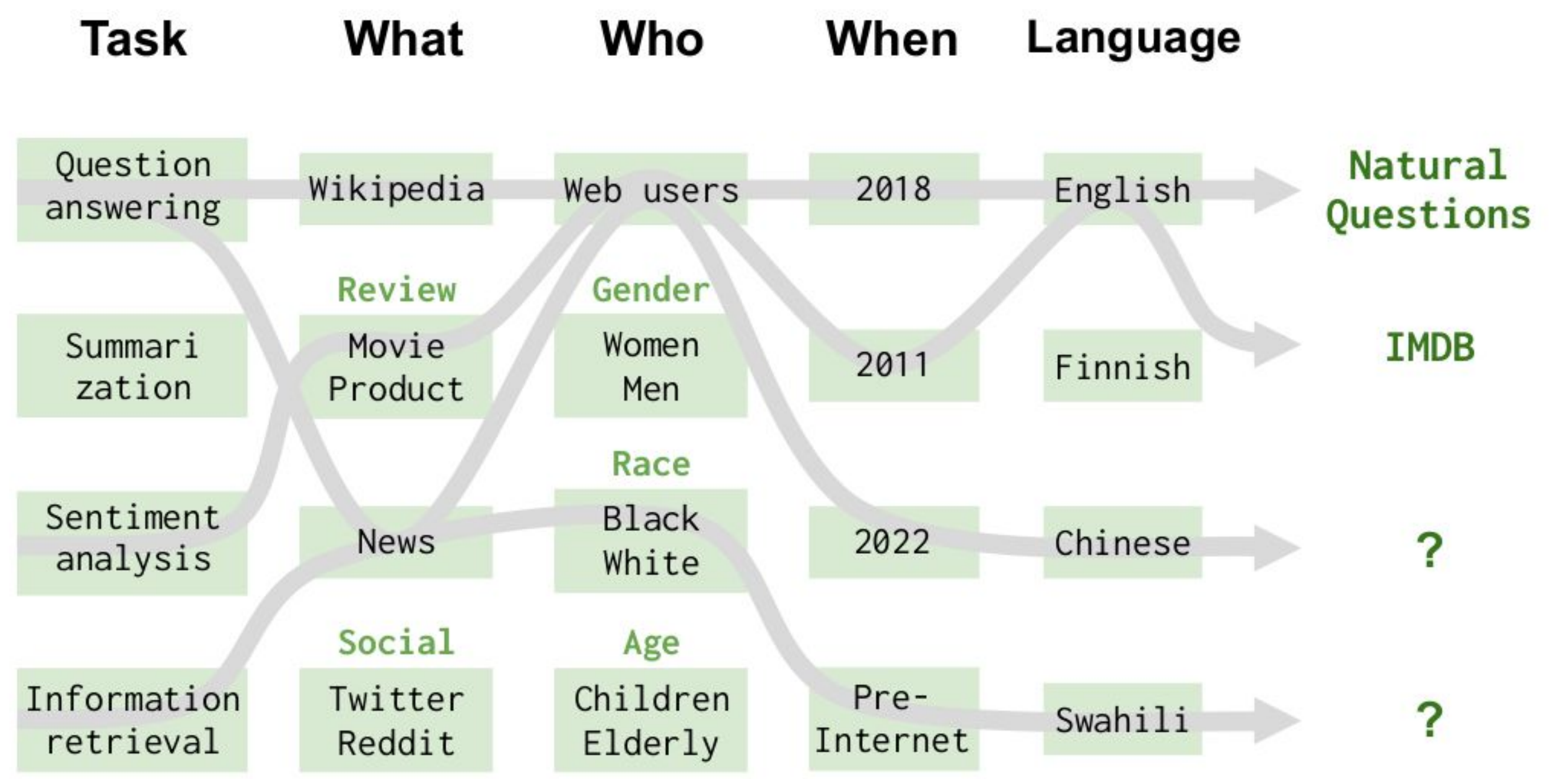

Figure 2: HELM Scenario Evaluation Scenarios

- HELM also defines standardized scenarios to test all models on, across a variety of measurements. This was developed alongside a variety of competing large models (OpenAI, Google, Meta, and others) to ensure neutral unbiased results.

Metrics

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

Figure 3: HELM Evaluation Metric Categories

MOTIVATIONS

- HELM aims to be a standardized objective metric for all language models. To accomplish this, its methodology is fully disclosed and public.
- We hypothesize that having an open benchmark platform poses challenges, as a malicious actor could manipulate the model's internals by knowing the dataset used for evaluation.
- This project aims to test this hypothesis using the HELM platforms by constructing fake models that may secure top scores in the benchmarks but are incapable of simple reasoning tasks that an ideal large language model should handle.

METHODOLOGY

- For this project, we used the Facebook BART-base model, developed by Facebook AI Research. BART is a transformer-based sequence-to-sequence model that excels in text generation, summarization, and comprehension, making it a solid foundation for our benchmark experiments. We fine-tuned two separate versions of BART using two datasets:
- OpenBookQA (by Mihaylov et al.) is a multiple-choice question dataset that tests basic science knowledge and elementary reasoning using a small "open book" of facts.
- LegalBench (from the Hazy Research group at Stanford) is a collection of legal tasks designed to evaluate a model's ability to handle complex legal reasoning, argument comparison, and statutory interpretation.
- This project aims to test this hypothesis using the HELM dataset by constructing light models that may secure top scores in the benchmarks but are incapable of simple reasoning tasks that an ideal large language model should handle.

EVALUATION RESULTS

Model trained on OpenbookQA and evaluated on other scenarios

Scenario	Metric Name	Data
MedQA	Quasi-Exact Match	0.15
NarrativeQA	F1 Score	0.05
WMT	BLEU-4	0.20
GSM8k	Final Number Exact Match	0.00
Math	Equivalent (CoT)	0.00
NQ (Open Book)	F1 Score	0.02
NQ (Closed Book)	F1 Score	0.02
MMLU	Exact Match	0.15
LegalBench	Quasi-Exact Match	0.00
OpenBookQA	Exact Match	0.99

Model trained on LegalBench and evaluated on other scenarios

Scenario	Metric Name	Data
MedQA	Quasi-Exact Match	0.00
NarrativeQA	F1 Score	0.00
WMT	BLEU-4	0.00
GSM8k	Final Number Exact Match	0.00
Math	Equivalent (CoT)	0.00
NQ (Open Book)	F1 Score	0.00
NQ (Closed Book)	F1 Score	0.00
MMLU	Exact Match	0.00
OpenBookQA	Exact Match	0.00
LegalBench (Control)	Quasi-Exact Match	1.00

DISCUSSION & CONCLUSION

- Our experimentation has shown that, by using open-source datasets and evaluation, an incompetent model can still receive high scores despite being unable to perform other basic tasks that would be expected out of an ideal large language model in practice.
- The open-source nature of the HELM benchmarking process is vital to ensure it produces trustworthy, fair, and scientifically objective results. However, it is still vulnerable to manipulation by bad actors looking to obfuscate inferior models. Investors and researchers should be wary to not take limited model evaluation data at face value, and perform their own tests.

REFERENCES

- Stanford Center for Research on Foundation Models. *HELM Lite: Holistic Evaluation of Language Models*. Stanford University, https://crfm.stanford.edu/helm/lite/latest
- Facebook AI. *facebook/bart-base*. Hugging Face, https://huggingface.co/facebook/bart-base
- Hazy Research. *LegalBench: A Benchmark for Legal Reasoning in Language Models*. Stanford University, https://hazyresearch.stanford.edu/legalbench/
- Mihaylov, Todor, et al. "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering." *arXiv*, 6 Sept. 2018, https://arxiv.org/abs/1809.02789

ACKNOWLEDGMENTS

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (CySER) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute.