Multifaceted Benchmarking[1]

Eduardo Neville George Golynskyi Tetra Jones

Organizers: Esben Kran, Jason Hoelscher-Obermaier, Fazl Barez, Marius Hobbhahn

Abstract

Currently, many language models are evaluated across a narrow range of benchmarks for making ethical judgments, giving limited insight into how these benchmarks compare to each other, how scale influences them, and whether there are biases in the language models or benchmarks that influence their performance. We introduce an application that systematically tests LLMs across diverse ethical benchmarks (ETHICS and MACHIAVELLI) as well as more objective benchmarks (MMLU, HellaSwag and a Theory of Mind benchmark), aiming to provide a more comprehensive assessment of their performance.

Keywords: Evals, AI security, governance

1. Introduction

We hypothesize that large language models are typically evaluated on a limited set of ethical benchmarks, potentially leading to a biased understanding of their capabilities and limitations. To address this limitation, we propose the development of an application that systematically tests LLMs across a spectrum of benchmarks, each representing distinct ethical considerations. We predict that exposing LLMs to a more diverse range of ethical challenges will provide a more comprehensive assessment of their performance, revealing potential biases and shortcomings that might be overlooked in a single-benchmark evaluation.

[1] Research conducted at The Model Evaluations Hackathon, 2023 (see https://alignmentjam.com/jam/evaluations)

Furthermore, we want to investigate the order in which ethical prediction and reasoning skills appear in language models; performance across the MMLU dataset shows that GPT-3 does better at college medicine and college mathematics than elementary mathematics, unlike the order in which humans learn (Hendrycks et al., 2021), and similar differences from human performance might exist for predicting ethical judgments.

2. Methods

We include the HellaSwag and MMLU benchmarks in our battery as commonly used evaluations to judge the reasoning abilities of language models. (Hendrycks et al., 2021; Zellers et al., 2019). Scores for open source Hugging Face models were available on the Open LLM Leaderboard (Beeching et al., 2023); OpenAI models were evaluated using OpenAI evals (*OpenAI Evals*, 2023/2023).

For ethical judgments explicitly, we use the ETHICS benchmark and the MACHIAVELLI benchmark (Pan et al., 2023). For the ETHICS benchmark, we use few-shot prompting for all language models, adapted from the approach used in (Hendrycks et al., 2023) for GPT-3, in order to enable like-for-like comparison between models for which the log probabilities are available (e.g. Mistral) and for which they are not (e.g. ChatGPT). The prompt for the virtue, deontology, justice and evaluations is a list of examples and their categories; the prompt for the utilitarianism questions is as in Hendrycks 2023.

We include a Theory of Mind benchmark from (gcmac & Nathan, n.d.) to compare ability to model humans against ability to predict ethical judgements.

These benchmarks are run using a script that systematically evaluates language models[2]; where needed, the codebases of the relevant benchmarks were adapted to handle both OpenAI and HuggingFace transformer models.

3. Results

The following is a table of the performance of models we have considered against the benchmarks we have run.

Model	ETHICS (Justice)	ETHICS (Deontology)	ETHICS (Virtue)	ETHICS (Utilitarianism)	ETHICS (Commonsense)	ETHICS (Justice Hard)	ETHICS (Deontology Hard)	ETHICS (Virtue Hard)	ETHICS (Utilitarianism Hard)	ETHICS (Commonsense Hard)	Theory of Mind	hellaswag	mmlu
gpt-3.5-turbo	0.76	0.74	0.76	0.95	0.63	0.75	0.58	0.87	0.95	0.51	0.75	0.79	0.6545
gpt-4-turbo	0.94	0.82	0.88	0.98	0.86	0.94	0.72	0.82	1	0.66	0.9	0.93	0.778
gpt2	-	0.5	-	-	0	-	-	-	-	-	-	-	0.324
mistralai/Mistral-7B-v0.1	-	-	-	-	-	-	-	-	-	-	-	0.8331	0.6416
teknium/OpenHermes-2.5-Mistral-7B	-	-	-	-	-	-	-	-		-	-	0.8424	0.6373

[2] Code available at https://github.com/EduardoNeville/Multifaceted-Benchmarking

4. Discussion and Conclusion

From our current data, there is not much of a conclusion to qualitatively draw; GPT-4-Turbo is predictably better than GPT-3.5-Turbo across every domain, with the exception of the hard virtue ethics dataset of ETHICS which we have not investigated enough to robustly conclude anything about.

The scripts we have written are able to run the relevant evaluations with weights from HuggingFace using the transformers library (Wolf et al., 2020), though during the time of the hackathon we were unable to run any models of nontrivial size due to compute limits (we attempted but GPT-2 was unable to achieve nonrandom performance on benchmark given); further work would involve running these evals against those models.

Some of our work is likely reimplementing features of EleutherAI's evaluation harness (Gao et al., 2021); we could investigate using that platform in whole or part to run more evaluations.

5. References

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., & Wolf, T. (2023). *Open LLM Leaderboard*. Hugging Face. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., Phang, J., Reynolds, L., Tang, E., Thite, A., Wang, B., Wang, K., & Zou, A. (2021). *A framework for few-shot language model evaluation* (v0.0.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5371628

gcmac, & Nathan. (n.d.). *Evaluating GPT-4 Theory of Mind Capabilities*. Retrieved November 27, 2023, from https://www.lesswrong.com/posts/Ce82o8mbBfH9N3Jes/evaluating-gpt-4-theory-of-mind-capabilities

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2023). *Aligning AI With Shared Human Values* (arXiv:2008.02275).

arXiv. https://doi.org/10.48550/arXiv.2008.02275

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (arXiv:2009.03300). arXiv. http://arxiv.org/abs/2009.03300

OpenAI Evals. (2023). [Python]. OpenAI. https://github.com/openai/evals (Original work published 2023)

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., & Hendrycks, D. (2023). *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark* (arXiv:2304.03279). arXiv. https://doi.org/10.48550/arXiv.2304.03279

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing* (arXiv:1910.03771). arXiv. https://doi.org/10.48550/arXiv.1910.03771

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *HellaSwag: Can a Machine Really Finish Your Sentence?* (arXiv:1905.07830). arXiv. https://doi.org/10.48550/arXiv.1905.07830