**Quantifying Bias Propagation in Domain-Adapted Foundation Models**

**PI: Dr. Bill Howe, Associate Professor, University of Washington**

Few now dispute that foundation models will reshape human society, with transformational effects expected in domains ranging from scientific inquiry to business administration to mass communication [3]. However, as these models are adapted and specialized for mainstream human social and economic activities, the risks of unanticipated biases increases. Consider, for example, the findings of a recent study on pretrained biomedical language models trained on the widely used MIMIC-III dataset, which found that valenced text descriptions and some medical conditions predicted biased inferences of race in the fine-tuned language model [1]. While NLP has reckoned with bias in foundation models for years [11], the potential harms are higher than ever as an ecosystem of applications across all sectors inherit the biases of relatively few models trained on undocumented, weakly curated datasets.

A diagram of a diagram of a health care system

Description automatically generated with medium confidence

Figure 1: Adapting the LLAMA-2 foundation model to consequential domains such as biomedical Q&A, occupation classification, and hate speech detection allows for assessment of performance and fairness in models fine-tuned using the latest techniques.

While most analysis focuses on the largest models, emerging evidence suggests that, under some circumstances, smaller pretrained models can be fine-tuned and specialized to a domain such that they are competitive with much larger models [5]. Especially encouraging results have been obtained from methods such as low-rank adaptation, which allow for models to be efficiently adapted on consumer-grade hardware using small tunable weight matrices [2]. ***However, the tradeoffs between smaller models and foundation models in bias-sensitive domains are not well understood.***

**Scope of work:** We will evaluate data curation as a means to control bias in domain-adapted foundation models. Existing work (including our own, e.g., [13]) has generally been limited to observational studies of bias in foundation models, as the training data and procedures are often undocumented, let alone exposed for experimentation. The AI responsibility community has increasingly recommended data curation to mitigate harms [15], but the necessary comparative experiments to assess this approach are prohibitively expensive for the largest models. The availability of LLAMA-2-7B [12] affords a controlled study at a feasible scale using *deterministically adjusted levels of training data bias*: We will compare models on representative tasks, adjusting the training data to simulate the effects of data curation, measuring performance and bias. Specifically, we will fine-tune LLAMA-2-7B over varying training data in the following scenarios:

1. **Hiring, via the Hybrid Hiring Dataset**. The Hybrid Hiring dataset was used in a human-AI collaboration study which found that more biased language technologies (such as the Bag-of-Words model employed in the study) resulted in more biased human-AI collaborative decisions [10]. The dataset consists of a web scrape of online biography webpages, and the NLP task is occupation prediction [10]. For this domain, we evaluate the fine-tuned models for their accuracy in occupation prediction, and for the gender bias they exhibit, quantified as differential true positive rate.
2. **Biomedical Question Answering, via the PubMedQA benchmark and the BBQ Bias Question Answering Task**. The widely used PubMedQA benchmark evaluates the ability of large language models to answer biomedical questions [6], and more closely aligns with a typical use case of instruction-tuned models in information seeking and retrieval. We will fine-tune LLAMA-2 on the PubMedQA dataset and related tasks such as MIMIC-IV [8] to adapt it to the biomedical domain. Fairness is assessed using the BBQ bias question answering tasks [9], which quantifies social biases in both underspecified contexts and disambiguated contexts.
3. **Hate Speech Classification, via TOXIGEN**. Language models are already widely used for online hate speech detection, but often struggle with *implicit* hate speech detection, wherein discriminatory messages that do not include charged words may be missed while lexically charged messages may be flagged. This phenomenon often results in biases affecting text written in dialect [4]. The TOXIGEN dataset and benchmark formalize the accurate detection of implicit hate speech detection [4]. The proposed research would fine-tune LLAMA-2 on TOXIGEN and evaluate accuracy of the model for classifying implicit hate speech, and the fairness impacts for minoritized groups.

**Experimental Design:** We will fine-tune on a random sample of the target dataset N=25 times to better understand the distribution of model performance and fairness on the task. We will then curate the fine-tuning dataset such that the sensitive attribute is paired with a target category at certain levels. For example, in the Hybrid Hiring dataset, we will set the proportion of female vs. male biographies for a profession like surgeon to 10/90, 20/80 … 80/20, 90/10, and fine-tune N=25 models on subsets of the dataset such that this proportion of male to female is maintained. At each level we will conduct paired-samples t-tests to assess whether the performance and fairness of models fine-tuned on the curated dataset differs from that of models fine-tuned on the uncurated dataset, taking Cohen’s *d* where the test is significant to obtain an effect size quantifying the magnitude of the difference. We will also use linear models to measure the relationship between the curated level of the sensitive attribute and the dependent variables of performance and fairness, with post-hoc analysis performed as appropriate where the null hypothesis is rejected.

**Use of resources:** The resources provided via by the Accelerate Foundation Models Research program will be used to fine-tune the LLAMA-2 language model and to gather new data as appropriate.

**Expected outcomes:** We expect to submit a research paper to a well-regarded AI fairness or NLP conference such as ACM FAccT or ACL, and to publicly release all trained models and ancillary data associated with publication. This work will provide a rigorous experimental design to assess the tradeoffs of data and inform the design of foundation models themselves, where controlled ablations are prohibitively expensive. In addition to the knowledge disseminated via publication, the release of a set of pre-trained, bias-adjusted models simplify downstream investigations, including the use of low-memory techniques such as Low-Rank Adaptation and Prompt Tuning for adapting models on consumer-grade hardware. Overall, this project aims to use Azure resources and LLAMA-2 to rigorously assess commonly proposed (yet rarely studied) data-level interventions to control bias mitigation and produce practical recommendations for the safe use of Foundation Models.

**References**

[1] Adam, Hammaad, et al. "Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022.

[2] Ding, Ning, et al. "Parameter-efficient fine-tuning of large-scale pre-trained language models." *Nature Machine Intelligence* 5.3 (2023): 220-235.

[3] Eloundou, Tyna, et al. "Gpts are gpts: An early look at the labor market impact potential of large language models." *arXiv preprint arXiv:2303.10130* (2023).

[4] Hartvigsen, Thomas, et al. "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.

[5] Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*. 2021.

[6] Jin, Qiao, et al. "PubMedQA: A Dataset for Biomedical Research Question Answering." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

[7] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3.1 (2016): 1-9.

[8] Johnson, Alistair, et al. "Mimic-IV." *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)* (2020).

[9] Parrish, Alicia, et al. "BBQ: A hand-built bias benchmark for question answering." *Findings of the Association for Computational Linguistics: ACL 2022*. 2022.

[10] Peng, Andi, et al. "Investigations of performance and bias in human-AI teamwork in hiring." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. No. 11. 2022.

[11] Sheng, Emily, et al. "The Woman Worked as a Babysitter: On Biases in Language Generation." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

[12] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." *arXiv preprint arXiv:2307.09288* (2023).

# [13] Wolfe, Robert, Yang, Yiwei, Howe, Bill, Caliskan, Aylin “Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias.”[*FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*](https://dl.acm.org/doi/proceedings/10.1145/3593013)

# [14] Yan, A., & Howe, B. (2020). Fairness-Aware Demand Prediction for New Mobility. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01),. <https://doi.org/10.1609/aaai.v34i01.5458>

# [15] Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, Shmitchell, Shmargaret, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*