Generative AI Governance: Technological Monoculture, Market Structure and the Risk of Correlated Failures

Ramayya Krishnan[1], Prasanna Parasurama[2], Joao Sedoc[3], Arun Sundararajan[4]

Given the significant barriers to creating high-quality foundation models (cost of collection of training data, need for access to immense computing power), a small number of primarily closed-source foundation models are establishing leadership in the generative AI market. Applications based on these foundation models are being deployed by a number of firms across multiple sectors.

Responsible use of AI requires an understanding of the safety and reliability of the AI models and their use in applications of societal consequence. The advent and success of large language models (LLMs) has changed the AI architectures that are being deployed in organizational applications. Specifically, LLMs are developed and trained without a single downstream use case in mind. Fine-tuning or otherwise customizing these general purpose models creates an instance of an AI model suited to the needs of an application. This platform model is a departure from the purpose built AI models designed to meet the needs of particular use cases.

Granted, LLM architectures have many advantages that are common in platform-based approaches, most notably the economies of scale and scope that flow from being able to draw on pre-trained capabilities rather than building them from scratch, However,AI applications derived from these models can suffer from correlated errors and risks. These errors and risks may arise in myriad downstream applications, ranging from recruiting to healthcare provision.

Additionally, and perhaps more saliently, given the extent to which the training data sets of the **generative AI** foundation models overlap, the risks could be far more substantial than what might be suggested by a competitive analysis of market structure and market shares. Indeed, a recent study by Zou et al. (2023) demonstrates a simple class of suffix attacks that exploit a vulnerability in all current aligned LLMs to get them to produce content that their guardrails were designed to prevent.

Understanding correlated risks is a topic that has not been extensively studied in the literature and is critical to the responsible use of AI in consequential application domains. This is a gap our study addresses.

---

[1] Heinz College of Information Systems and Public Policy, Block Center for Technology and Society, Carnegie Mellon University. rk2x@cmu.edu
[2] Goizueta Business School, Emory University. pparasurama@emory.edu
[3] Leonard N. Stern School of Business, New York University. jsedoc@nyu.edu
[4] Leonard N. Stern School of Business, New York University. digitalarun@nyu.edu

In this study, we analyze the relationship between the **diversity** in upstream foundation models and the risk of **correlated failures** and shared vulnerabilities in downstream applications**.** Such risks are similar to those generated by **monoculture** in farming settings (Power and Follett, 1987) wherein reliance on fewer seed strains can lead to shared crop vulnerabilities to pathogens and a higher risk of famine. They have also been highlighted in the digital context, most notably about vulnerabilities in information security (Birman and Schneider, 2009, Chen, Kataria and Krishnan, 2011). For example, the vast market share of the Windows operating system has for decades provided malicious agents with the incentive to invest effort to discover and exploit its information security vulnerabilities.

More recently, the risks of algorithmic monoculture have been raised for AI, largely in studying algorithmic screening of job applicants. In particular, Kleinberg and Raghavan (2021) analyze the case where firms that compete on hiring have a choice of using algorithmic hiring or manual processes and demonstrate that homogeneity in the algorithm used by the competing firms leads to a type of Braess' paradox: the introduction of a more accurate algorithm can drive the firms into a unique equilibrium that is worse for society than the one that was present before the algorithm existed. Bommasani et al. (2022) develop a simple mathematical formalism to measure systemic failure where the same individual is rejected by every firm that they apply to on account of the homogeneity of the resume processing algorithm in use. Their subsequent measurement experiments study the extent of correlation in outcomes depending on which adaptation method was used to adapt the foundation model.

Building on this stream of literature, we study the extent to which foundational large language models pose a systemic risk of correlated failures in a high-stakes setting: algorithmic screening of job applications. We consider a scenario in which multiple firms use the same foundational model to fine-tune a resume-screening algorithm using their own data. We ask whether the use of the same foundational model contributes to correlated errors (false negatives and false positives) across firms – i.e., whether the same individual would be incorrectly rejected (false negative) or incorrectly selected (false positive) across firms.

We use applicant tracking system (ATS) data from 8 firms based in the U.S. The ATS tracks all details of the firm's job postings (job title, department, job description), job applications (candidate details, demographics, resume text), and the outcome of each application (whether the applicant received a callback). The data spans 2014-2018, containing 1.17M job applications for 6.6k job postings. Since we are interested in correlated errors across firms for a given individual, we identify the subset of individuals that applied to similar positions at multiple firms within the same time period in our dataset. This amounts to 25k individuals with 65k applications to 3.6k jobs across 8 firms.

In our initial set of baseline experiments, we prompt off-the-shelf LLaMA-2-7B and LLaMA-2-13B models, both foundational LLMs released by Meta, with the candidate's resume and the corresponding job description and ask whether the candidate should receive a callback. We compare these algorithmic predictions to the ground truth in our ATS data (whether the candidate received a callback), and estimate the level of correlated errors across firms. Our results show that the off-the-shelf model has almost no predictive power for this screening task, leading to uncorrelated errors across firms.

Subsequently, we use parameter efficient fine-tuning to create 8 different LLaMA-2-7B models, one for each firm, on the respective firm's hiring data, and study how the level of correlated errors changes with the model's predictive power. Our preliminary findings show that both standard machine learning models (e.g., logistic regression + tf-idf) and fine-tuned LLaMA-2-7B models have similar AUC. However, unlike baseline machine learning models, Llama-2 fine-tuned models show significant correlated false negatives between firms. ‘