

UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”

PHD SCHOOL

PhD program in: Economics and Finance

Cycle: 36

Disciplinary Field (code): SECS-P/07

**Essays in Economics of Artificial
Intelligence**

Advisor: Carlo Rasmus Schwarz

Co-Advisor: Francesco Decarolis

PhD Thesis by

Mahyar Habibi

ID number: 3144039

Year 2025

Abstract

This thesis embodies three chapters on the economics and applications of artificial intelligence (AI). The first chapter explores the economic underpinnings of open-source contributions in AI by for-profit companies, focusing on large language models (LLMs). Three main findings emerge: (1) LLMs align well with the R&D portfolios of diverse technologically advanced firms, (2) models developed by large technology companies are more likely to be open-sourced, and (3) open-sourcing advanced LLMs enhances research-related activities. A theoretical framework analyzes the factors influencing a firm's decision to open-source, suggesting an inverted-U-shaped relationship between open-sourcing propensity and the firm's share of LLM-compatible applications. The second chapter addresses the debate on the moderation of toxic speech on social media and its impact on the plurality of online discourse. A new methodology is proposed and validated to measure plurality based on the semantic variance of online content, using text embeddings from computational linguistics. Applying this measure to a dataset of 10 million US Tweets, it is found that removing toxic content reduces the plurality of online discourse. Crucially, the reduction in plurality is attributed not to the toxic language itself, but to the removal of meaningful content. The third chapter proposes a novel method for estimating biases at the micro-level in contexts with multiple bilateral interactions, where individual preferences and correlated characteristics complicate analysis. The method employs Collaborative Filtering in an 'honest' design to extract preferences and characteristics, separating self-induced outcomes from the constructed embeddings of interacting units.

Contents

1	Open Sourcing GPTs	7
1.1	Introduction	7
1.2	Context; LLMs in a Nutshell	14
1.3	Data	17
1.4	Measuring the Scope of Application of LLMs	20
1.4.1	Crafting the Latent Technology Space	20
1.4.2	Firms' Compatibility with LLMs and open-source Contributions	23
1.5	Empirical Analysis	27
1.5.1	Model Quality and Open-Sourcing Decisions	27
1.5.2	Open Source as an R&D Catalyst	31
1.6	Theoretical Analysis	41
1.6.1	Environment	41
1.6.2	Open Source Decision	44
1.6.3	Results	48
1.7	Conclusion	53
	Appendix	56
2	Content Moderator Dilemma	77
2.1	Introduction	77

2.2	Data and Methods	81
2.2.1	Representative US Twitter Data	81
2.2.2	Measuring Online Plurality	82
2.3	Results	85
2.3.1	Removal of Toxic Content and the Plurality of Online Discourse . .	85
2.3.2	Controlling for Toxicity	88
2.3.3	What Drives Reduction in Plurality	89
2.3.4	Alternative Forms of Content Moderation	91
2.4	Conclusion	92
	Appendix	94
3	Separating Biases from Preferences	117
3.1	Introduction	117
3.2	Conceptual Framework and Methodology	121
3.3	Simulated Example	126
3.4	Empirical Example	131
3.4.1	Data Description	131
3.4.2	Estimation	132
3.5	Conclusion	137

Acknowledgements

I would like to express my deepest gratitude to my advisors, Professors Francesco Decarolis, Andrea Fosfuri, Avi Goldfarb, and Carlo Schwarz, for their invaluable guidance, support, and encouragement throughout this journey. Your insights and expertise have been instrumental in shaping this work.

A special thanks to my co-authors, Prof. Dirk Hovy, Dr. Zahra Khanalizadeh, and Dr. Negar Ziaeeian, whose collaboration and shared knowledge have significantly contributed to this thesis. I am also grateful to the coordinators of the PhD program, Professors Mariano Massimiliano Croce and Marco Ottaviani, for their continuous support and encouragement.

To my family, thank you for your unwavering support, love, and patience. Your belief in me has been my constant source of strength. Finally, I dedicate this work to the memory of my father. He was with me at the beginning of this journey but is no longer with us to see its completion. His memory continues to inspire me every day.

Introduction

The rapid progress of artificial intelligence (AI) systems and machine learning (ML) methods has opened numerous avenues for research within economics and the broader social sciences. This thesis comprises three chapters that explore the economics of AI and the application of AI and ML methods in social science research.

The first chapter of this thesis explores the influential role of large for-profit technology companies in the development and open-sourcing of artificial intelligence, particularly large language models (LLMs). It examines the factors driving these companies' decisions to share AI breakthroughs openly or keep them proprietary. The empirical analysis reveals three key findings: the LLMs technology is compatible with the R&D portfolios of many technologically differentiated firms, models developed by large tech companies are more likely to be open-sourced, and open-sourcing advanced LLMs boosts research-related activities. A theoretical framework is proposed, framing the open-sourcing decision as a trade-off between accelerating growth and securing financial returns, predicting a higher tendency to open-source when a firm's LLM moderately outperforms existing alternatives, and indicating an inverted-U-shaped relationship between the propensity to open-source and the firm's share of LLM-compatible applications.

The second chapter addresses the growing concern of hateful and inflammatory content online and the subsequent rise in content moderation efforts by online platforms. While

these efforts aim to curb real-life violence linked to such content, they have also sparked debates on free speech and the plurality of online discourse, presenting a dilemma for platforms. This chapter proposes and validates a methodology to measure content plurality using text embeddings from computational linguistics. Analyzing a dataset of 10 million US Tweets, it finds that removing toxic content reduces online content plurality, regardless of the embedding model, toxicity metric, or variance measure. The reduction in plurality is not due to the toxic language itself but the removal of meaningful content. The chapter suggests an alternative moderation approach using LLMs to rephrase toxic content, demonstrating that this method reduces toxicity without diminishing content plurality.

The third chapter introduces a method for examining discriminatory behavior at the individual level, addressing the limitations of aggregate-level studies. Estimating discrimination at the micro level is important for regulatory efforts to identify and mitigate different types of discrimination, especially in the labor market. The method consists of two stages: the first stage employs Collaborative Filtering with an “honest” design to extract preferences and characteristics, separating these from self-induced outcomes. The second stage uses a Double Machine Learning estimator to identify biases at the unit level. Applied to a dataset of nearly 150,000 film ratings by professional critics, the study uncovers that approximately 5% of critics exhibit a significant bias in favor of films directed by women after accounting for personal preferences and film characteristics. In contrast, a “naive” approach that ignores these factors suggests a much higher prevalence of bias.

Chapter 1

Open Sourcing GPTs

Mahyar Habibi

1.1 Introduction

Open source contributions have significantly shaped the growth of artificial intelligence, machine learning, and more recently large language models (LLMs). Interestingly, large for-profit technology companies have played crucial and at times dual roles in this rapidly evolving landscape. On one hand, these companies have made notable contributions to the open source ecosystem by sharing scientific breakthroughs such as Transformer architecture and open-sourcing advanced software like TensorFlow, PyTorch, and LLaMA. The extent and impact of their contributions over the past decade arguably surpass those made by the most prolific academic institutions (Ahmed et al., 2023). On the other hand, following recent breakthroughs in LLM capabilities, some major technology firms have revised their stance toward the open source ecosystem. They now restrict and monetize access to their LLMs while expressing concerns about the dangers of open-sourcing advanced models (e.g., Post, 2023; WSJ, 2024).

This paper argues that open-sourcing advanced AI models like LLMs presents profit-maximizing firms with a strategic trade-off between accelerating technological growth

and securing immediate financial returns. The key takeaway is that firms are most likely to open source multi-purpose software such as LLMs when they own a significant but not excessive share of compatible applications. Small firms with few compatible applications prefer a closed strategy for immediate revenue, while firms dominating compatible applications find open source community contributions insignificant compared to their internal resources. However, for technologies with wide-ranging use cases like LLMs, even Big Tech giants own a modest share of compatible applications, potentially finding the benefits of open sourcing outweigh the costs. Meta CEO Mark Zuckerberg’s remarks on Generative AI and the company’s LLaMA open sourcing strategy align with this argument. Zuckerberg stated, “In the last year, we have seen some really incredible breakthroughs — qualitative breakthroughs — on generative AI and that gives us the opportunity to now go take that technology, push it forward, and *build it into every single one of our products*,” and while he does not expect LLaMA to generate “a large amount of revenue in the near term, but over the long term, hopefully that can be something” (CNBC, 2023a,b). This insight contributes to our understanding of the economics of open sourcing in AI, highlighting how the properties of AI as a potential general-purpose technology influence firms’ strategic decisions and, consequently, the AI development trajectory.

The analysis proceeds in four parts. The first part examines the compatibility of LLMs in the R&D process of innovating firms. For this part, I use patent data and propose a novel strategy to examine compatibility of firms’ R&D process with LLMs. I find that LLMs are compatible with R&D portfolios of a large set of technologically diverse firms, implying a broad range of industrial applications for this technology.

In the second part, I examine the relationship between the quality of the models and the open-sourcing strategy of the developers. Using data on major model releases and their performance on a widely used benchmark, I find that a 10-point increase in quality (on a 100-point scale) over the existing state-of-the-art open source model is associated

with a 10-11 percentage point decrease in the likelihood of the model being open sourced. Furthermore, for-profit organizations are, on average, 14-18% less likely to open source a model. However, the analysis suggests that Big Tech companies, *ceteris paribus*, are 20% more likely to open source a model than other for-profit organizations. In the third part of the analysis, I combine AI/ML-related publication records with GitHub data and document a significant increase in research-related activities among LLM researchers following the open source release of LLaMA, an advanced LLM developed by Meta. This finding implies that open-sourcing advanced software can stimulate related R&D efforts.

Motivated by these findings, I propose a theoretical framework in the final part of the analysis to examine the decision-making process of for-profit firms in developing and open-sourcing a new LLM. In the theoretical analysis, LLMs are framed as a potential general-purpose technology (GPT), capable of boosting profits in various applications. The model is structured as a two-stage decision-making process. Initially, a firm assesses the quality of the existing open source model to decide whether to develop a new LLM. In the second stage, the firm decides how to optimally allocate computational resources for integrating the model into its applications. Should the firm opt to develop a new LLM, it then faces a choice: permanently open source the model or keep it proprietary for an additional period. This decision presents a strategic trade-off: stimulate software growth and R&D efforts through open-sourcing or secure immediate profits by licensing. By open-sourcing, a firm leverages external contributions to enhance the model, accelerating its growth and integrating it more effectively with applications to boost profits. Alternatively, a closed-source release enables immediate revenue through API sales to external software producers, at the expense of missed community contributions.

The theoretical analysis generates several key predictions aligned with empirical findings. It suggests that the open-sourcing decision depends strongly on the quality lead over alternative open source models, with larger leads favoring closed-source strategies.

The analysis predicts an inverted-U shaped relationship between firm size and open-sourcing tendency, reflecting varying benefits from accelerated growth at different scales of LLM-compatible applications. Additionally, while both small and large firms may find developing new LLMs profitable when existing open source quality is modest, only larger firms are likely to do so when high-quality open source alternatives exist. Moreover, the model reveals nuanced effects of open source ecosystem efficiency. In a strong ecosystem, open-sourcing a marginally superior model may be beneficial. However, as the quality gap widens, open-sourcing becomes less attractive and the firm may have incentives to limit the efficiency of the open source ecosystem, thereby slowing the progress of open source rivals. This insight is particularly relevant given recent calls from major tech companies to regulate open source releases of advanced models (e.g., Post, 2023; Business-Insider, 2023).

AI is not the only field that saw significant contributions from for-profit companies to its open source ecosystem. Much of the infrastructure of Internet rests on foundations that were open sourced by for-profit firms, as well as operating systems for personal computers (Linux) and mobile devices (Android). Consequently, there's an extensive literature on the economics of open source software. This literature typically falls into two, sometimes overlapping categories. The predominant category examines programmers' motivations for contributing to open source projects. Though these incentives are crucial to the open source ecosystem of AI, my study does not delve into the individuals' incentives for contributing to open sourced AI projects. Instead, I focus on modeling the open-sourcing decisions of firms where a functioning open source community exists. For those interested in the incentives of open source contributors, Lerner and Tirole (2002) offers a comprehensive introduction to this area.

The second stream of literature on open source software, examines why firms choose to open source their proprietary software. This phenomenon extends beyond AI, with a

history of strategic open-sourcing decisions in various for-profit sectors. Existing research predominantly identifies the attraction of users to complementary proprietary products as a key driver for open sourcing (e.g., Lerner and Tirole, 2002; von Hippel and von Krogh, 2003; Lerner et al., 2006; Fosfuri et al., 2008). However, other motivations are also discussed. Henkel (2004) discusses standard-setting and signaling technical prowess, while Economides and Katsamakas (2006) considers open-sourcing as a platform strategy to benefit from proprietary applications built upon it. Gambardella and von Hippel (2018) points out that downstream firms may collaborate on open source alternatives to bypass upstream suppliers, and Nagle (2018) highlights the learning benefits firms gain from crowd feedback in open source projects. The theoretical framework in this study draws parallels to the competition between for-profit and non-profit entities in operating systems in Casadesus-Masanell and Ghemawat (2006), where the focus is on demand-side learning.

I make two contributions to this strand of literature. Firstly, despite being frequently discussed in the literature (e.g., Lerner and Tirole, 2002; Lerner et al., 2006), empirical evidence concerning the impact of open source software on encouraging research activities in a causal framework is rare. To the best of my knowledge, this study is the first to document empirical evidence concerning the potential effects of open source software on stimulating research activities. Nagle (2019) studies the impact of using open source software on firms' productivity and finds a positive and significant impact on the subset of firms with an ecosystem of complements. However, I am not aware of a study that directly investigates the impact of open source on research activity within a causal framework. Secondly, this study departs from existing literature by treating software not just as a product but as an enabling technology with applications across various sectors, generating nuanced insights into strategic development and open sourcing decisions not fully captured by existing frameworks.

Instances of inventors sharing technological advancements openly are rare, but not exclusive to AI. This phenomenon, termed “collective invention” by Allen (1983), was observed in 19th-century iron-making in Britain’s Cleveland district, where companies freely exchanged blast furnace design improvements. Similar patterns emerged in post-1800 steam engine enhancements (Nuvolari, 2004) and the flat panel display industry’s evolution (Spencer, 2003). Osterloh and Rota (2007) further suggest that open source software development is a modern embodiment of this collective invention concept. The open source ecosystem in AI and LLMs shares similarities and differences with historical collective invention cases. A common thread is the reliance on experimental trial and error, where shared experiences significantly enhance learning opportunities. However, in contrast to the AI ecosystem, where large tech companies play a pivotal role, historical episodes of collective invention often featured smaller firms with limited R&D resources. This study proposes that open source contribution of tech firms in AI is attributable to the broad applicability of AI, extending beyond the scope of any single firm. Consequently, the opportunity for each major tech firm to leverage community resources for the rapid advancement of their models remains substantial.

This study also relates to recent work examining the changing dynamics between industry and academic research. Arora et al. (2020) and Arora et al. (2021) document how corporate labs have shifted away from basic research towards development activities, potentially hindering the emergence of general-purpose technologies. They argue that firms’ scientific research decisions are shaped by a trade-off between internal benefits and spillover costs to rivals, suggesting this dynamic has contributed to declining corporate research investment. My analysis suggests that the tension between knowledge spillovers to rivals and appropriability may be partially mitigated when the technology’s application domain is sufficiently expansive and firms can protect their competitive advantage through downstream specialization, offering a new angle to understand open sourcing advanced

AI systems by big tech companies.

This paper also contributes to the rapidly growing field of the economics of AI. A growing strand of literature focuses on AI and more recently LLMs characteristics as a general-purpose technology (e.g., Brynjolfsson et al., 2018; Cockburn et al., 2018; Agrawal et al., 2023a,b; Goldfarb et al., 2023; Eloundou et al., 2023). Beyond the analysis of AI as a GPT, Jacobides et al. (2021) and Ahmed et al. (2023) highlight the dominance of few Big Tech firms in terms of resources and influence on AI research. The role of open source in AI is further examined by Rock (2019), studying how open-sourcing TensorFlow by Google affected the market valuation of AI-focused companies. This study contributes to this literature by exploring how characteristics of LLMs, as a potential general-purpose technology, influence firms' decisions to open source their models, and consequently, the technology's development trajectory.

Lastly, the method proposed in this paper for obtaining latent technology representation of firms and technologies can contribute to the broader innovation literature interested in examining similarities and differences in R&D processes using patent data. Recently, there has been growing interest in using unsupervised NLP techniques to represent a firm's R&D portfolio within a latent vector space (e.g., Arts et al., 2021; Hain et al., 2022). However, the popularity of unsupervised techniques in AI/ML is primarily driven by the unavailability of enough labeled training data by domain experts (Hovy, 2022). This contrasts sharply with patent data, where patents are classified by domain experts into comprehensive and detailed patent classification systems. Although efforts to use patent classification systems to represent firms' technologies go as far back as Jaffe (1986), the challenges posed by the discrete and rigid structure of patent classification systems have encouraged researchers to adopt unsupervised techniques for these purposes. Inspired by classical NLP and ML techniques, I propose a flexible method that overcomes these challenges and creates a technology latent space using "gold standard" data without

relying on fully unsupervised techniques.

The remainder of this paper is structured as follows: Section 1.2 provides a brief overview of the ecosystem of LLMs. Section 1.3 describes the data. Section 1.4 introduces the method used to create the latent technology space and analyzes LLMs within the constructed technology landscape. Section 3.4 studies the open-sourcing decisions in the LLM ecosystem and examines impact of open-sourcing LLaMA on research activity of LLM-researchers. Section 1.6 introduces the theoretical framework and outlines its predictions, and Section 3.5 concludes the paper.

1.2 Context; LLMs in a Nutshell

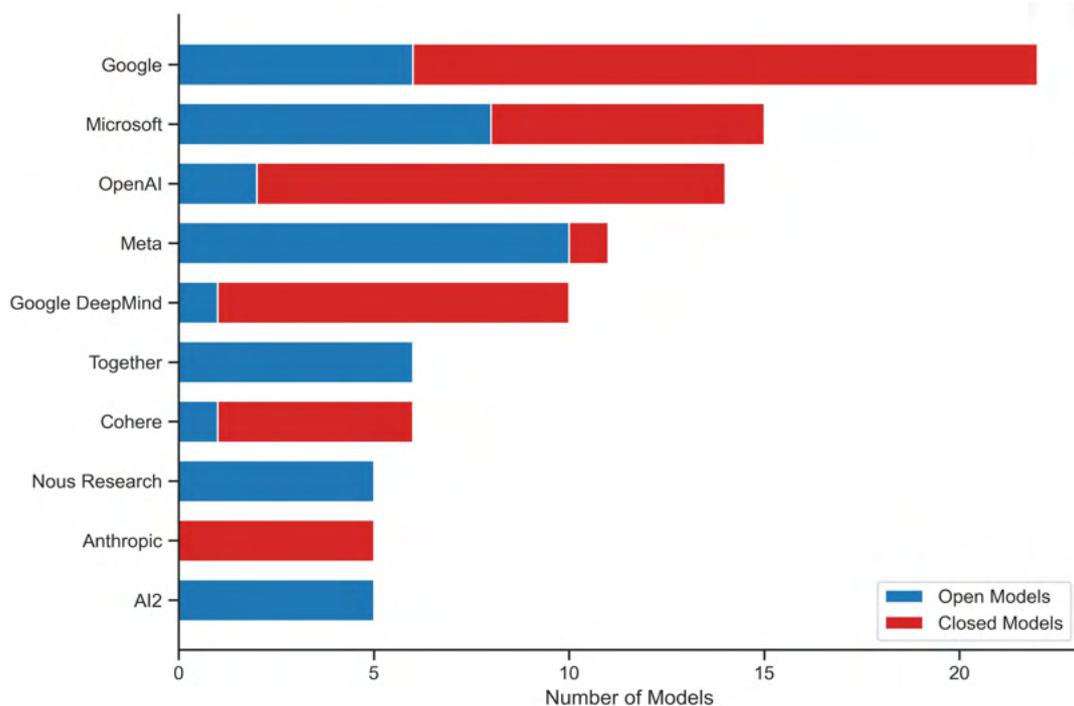
Although not clearly defined, Large Language Models (LLMs) can be broadly described as models based on artificial neural networks with billions of parameters, trained on a vast amount of text data in an unsupervised fashion, and capable of processing and often generating natural language data. In practice, however, LLMs mostly refer to models, often with tens of billions of parameters, built on the *Transformer* architecture proposed by Vaswani et al. (2017). The first generation of models now commonly recognized as LLMs, including BART, GPT-2, and T5, were released in 2019 (Lewis et al., 2019; Radford et al., 2019; Raffel et al., 2020). Since then, there has been a significant increase in the quality, quantity, and scale of such models. This section briefly describes the development process and the open source ecosystem of LLMs.

Large Language Models often undergo two main phases during their development: pre-training and fine-tuning. During pre-training, the model is exposed to a vast amount of text data (often tens of Terabytes) and is trained to predict the next word in a sentence given the previous words. This process requires massive computing resources (thousands of GPUs), can take months, and costs tens of millions of dollars for the state-of-the-

art models (NYT, 2023). The result is a versatile model capable of generating coherent text but often unable to provide desired responses for specific applications (e.g., chat bot). The purpose of fine-tuning is to adapt the pre-trained model to a specific task or domain and involves updating the parameters of the pre-trained model on a smaller, task-specific dataset (Howard and Ruder, 2018). It is noteworthy that fine-tuning, often done using relatively small datasets (e.g., 10-100K examples), incurs costs that are a fraction of the pre-training costs. The fine-tuning step can also involve additional steps such as reinforcement learning from human feedback (RLHF), which integrates human judgments directly into the fine-tuning process and allows models to learn preferences that are difficult to capture with traditional datasets or reward structures.

In the domain of LLMs, open-source is more nuanced than what is traditionally perceived as open source software (OSS). OSS can be loosely defined as software projects with published source code accompanied by a license allowing modification and redistribution (Lerner and Tirole, 2002). However, for LLMs, the training code is neither the only nor the most critical component of the software. The key components enabling practical use are the model parameters or weights. These weights can be made public without specifying the full training procedure or the model’s architecture. Additionally, datasets for fine-tuning the model can either be open-sourced or kept proprietary. Furthermore, there are stark differences among contributors within the open-source ecosystem of LLMs. A model’s performance, in terms of next-word prediction accuracy, depends largely on the model’s size, dataset volume, and computing resources dedicated to training (Kaplan et al., 2020). This so-called “scaling laws” of language models implies that developing state-of-the-art LLMs from scratch incurs significant costs, limiting the ability of many organizations to contribute new pre-trained models to the open-source community. However, when it comes to open-sourcing datasets, new training or inference methods, or releasing fine-tuned models, the open-source community is more diversified.

Figure 1.1 illustrates the number of open and closed models for ten leading organizations according to the ecosystem dataset from the Center for Research on Foundation Models (CRFM) at Stanford University¹. Google, OpenAI, Microsoft, and Meta lead in the number of models released. While most organizations have released both open and closed models, their strategies for open-sourcing vary significantly. Prominent AI startups such as OpenAI, Cohere, and Anthropic tend to keep their models primarily closed. Among the Big Tech companies, Meta has released more models publicly, and its LLaMA-series models are among the largest and most widely used open models. Google, on the other hand, employs a different strategy by keeping its flagship and larger models closed, while continuing to open-source smaller models.



Notes: The figure illustrates the number of open and closed LLMs by ten leading LLM developers in the ecosystem dataset of Center for Research on Foundation Models at Stanford University.

Figure 1.1: The Number of Open and Closed Models by Leading LLM Developers

¹This dataset includes only text-based models as well as multi-modal models such as text to image or text to audio.

1.3 Data

This study collects data from multiple sources for a comprehensive analysis at the firm and researcher levels within the open-source (OS) ecosystem of large language models (LLMs). Data were extracted from *Papers-with-Code*², *arXiv*, and *GitHub* to construct proxies for firms' contributions to the open-source community and activities of LLM researchers. Additionally, an analysis of patent application data filed with the US Patent and Trademark Office (USPTO) was conducted to evaluate the technology compatibility of firms and their engagement with Foundation Models. Below, a detailed description of each data source is provided.

- **Papers-with-Code** is a community-driven initiative led by the core team at Meta AI Research. It provides practitioners with free access to AI/ML research resources. This platform maintains up-to-date information on open-access AI/ML publications and tracks the presence of both official and unofficial code repositories associated with each paper. I retrieved the data in January 2024, focusing on papers that have an official repository on GitHub and were published on arXiv in 2019 or later. The resulting dataset contains more than 108 thousand publications.
- **arXiv**: Data on the initial publication dates, titles, and abstracts of papers were collected from arXiv. This analysis focused on papers published from 2019 onward, coinciding with the release of the first generation of LLMs such as GPT-2 and T5. Identifying papers related to LLMs was based on the analysis of its title and abstract, the methodology of which is detailed subsequently in this section.
- **GitHub**: I consider an open-access paper listed on Papers-with-Code with an official GitHub code repository as an open-source contribution.³ I extract two critical

²Paperswithcode.com

³The presence of an official code repository differentiates papers that contribute tangibly to the open-

pieces of information from GitHub: information on repository owners and data of contributors to these repositories.

- **USPTO:** This study employs patent data to identify firms that utilize generative language models in their R&D efforts, to assess the compatibility of a firm’s technology with LLMs, and to evaluate the breadth of applications firms aim to integrate with this technology. Considering LLMs’ status as an emerging technology, patent application data was preferred over granted patent data because the latter captures innovation activities only after a delay of at least a few years. The data were accessed through PatentsView.org in February 2024, providing information updated through December 31, 2023. The analysis is confined to utility patent applications filed by organizations with at least two applications during 2019-2023. The dataset comprises nearly 1.4 million applications from c.a. 61 thousand organizations.

Merging the Datasets

The Papers-with-Code dataset includes URL links to the respective arXiv pages and GitHub repositories, provided there is an associated page or repository for the paper. Linking organizations that own these repositories with those listed in patent application data is less straightforward. A primary challenge arises because the GitHub data often represent research groups within various institutions. Typically, details about these organizations are available on the organizations’ biography pages; however, this information is largely unstructured. To tackle this, I employ a language model to parse the information and identify profiles associated with commercial entities. After cleansing the names of applicants and repository owners, I merge the datasets using exact matching on cleaned names and unique-part matching for the remaining subset, subsequently removing false matches through manual inspection. I further analyze unmatched organizations with high

source ecosystem from those that only describe model performance across various benchmarks and tasks without any open-source contribution.

string similarity for potentially overlooked matches, adjusting the organization names in both datasets for initial match compatibility. Ultimately, approximately 180 organizations across the two datasets were successfully linked. For further details regarding the matching procedure and data parsing with the language model, see Appendix ??.

Identifying LLM-Related Papers

To identify papers related to LLMs, a narrow keyword search was deemed insufficient due to the rapidly evolving technical vocabulary in the field, which could either omit relevant papers or yield excessive false matches. To address this challenge, I fine-tuned an LLM classifier specifically for identifying LLM-related papers through a two-step process. Initially, two commercial LLMs, GPT-3.5 and Mixtral 8x7B, annotated a set of 20,000 out-of-sample papers⁴. Two separate models were employed to mitigate the reliance on a single model's classification outcome. The annotated dataset then served to fine-tune the pre-trained SciBERT language model (Beltagy et al., 2019) for this specialized task, achieving accuracy of 0.97 and an F1 score of 0.78 on a holdout sample. Finally, I used the fine-tuned model to identify LLM-related papers in the main sample.

Patent Applications

I leverage the Cooperative Patent Classification (CPC) system to identify firms integrating generative language models into their innovation activities. Additionally, I use applicant information to link organizations in the patent application data with those in the GitHub dataset, containing firms contributing to the open-source ecosystem of LLMs. Furthermore, I apply the CPC system to gauge firms' compatibility with LLMs and the scope of their R&D activities involving this technology, a process I will detail in Section 1.4.1.

⁴This sample originated from papers associated with unofficial code repositories on GitHub, in contrast to the main analysis focusing on papers with official repositories. To ensure reproducibility, the models' temperatures were set to zero.

1.4 Measuring the Scope of Application of LLMs

The primary objective in this section is to study the scope of industrial applications of LLMs through the lens of patent data. By analyzing technological differences among firms that could potentially leverage LLMs in their R&D processes, I aim to develop a better understanding of the environment in which open-sourcing decisions take place. To this end, I first outline the method I developed to assess the compatibility of firms' technologies with LLMs in a latent technology vector space. The analysis suggests a large number of firms in the patent application dataset have high compatibility with LLMs. Moreover, I find significant variation in the R&D portfolios of firms with potentially LLM-compatible technologies, suggesting a broad spectrum of industry applications for LLMs. I also examine data on the leading for-profit contributors to the open-source ecosystem for LLMs and find that the majority of LLM applications extend beyond the R&D scope of any single firm. These insights into the broad applicability of LLMs will form a cornerstone of the theoretical framework introduced in Section 1.6.

1.4.1 Crafting the Latent Technology Space

In this subsection, I present a novel methodology to create a latent technology space by leveraging the richness of patent classification systems and the flexibility of ML techniques. The goal is to represent each firm's overall R&D portfolio and LLMs in a high-dimensional vector space. The distance between a firm's vector and LLMs' vector will be used as a proxy for the firm's R&D compatibility with LLMs technology.

Mapping firms' technological positions in a vector space through patent classification has been a longstanding practice (e.g., Jaffe, 1986). Nevertheless, the discrete and hierarchical nature of the patent classification constrains the capability of traditional methodologies to capture nuanced technological profiles of firms. For instance, Jaffe's seminal method

employed an “ad hoc” categorization of over 300 patent classes into 49 groups, a schema likely too coarse to discern subtle technological distinctions. To address these shortcomings, researchers have employed NLP techniques to construct more refined vector spaces from patent texts (e.g., Arts et al., 2021; Hain et al., 2022).

However, using NLP techniques to represent technologies presents at least two major drawbacks. First, unsupervised approaches fall short of expert labeled data in capturing high quality information in complex tasks (Hovy, 2022). This issue is exacerbated with new technologies, where there might not be sufficient training data available about the new technology to enable the models to create an accurate representation of the technology in an unsupervised manner⁵. The second drawback concerns resources and efficiency. Even basic NLP techniques, create challenges for researchers when applied to a large volume of patent data (e.g., Kelly et al., 2021).

The proposed method is detailed in Algorithm 1. Initially, the method constructs a rich representation of patents by leveraging the hierarchical structure in the Cooperative Patent Classification (CPC) system⁶. For example, consider CPC code G06F40, which denotes the handling of natural language data. This code breaks down into section G, denoting Physics; subsection G06, specifying Computing, Calculating, or Counting; and class G06F, representing Electric Digital Data Processing. Typically, a patent is classified with multiple such codes. The method converts this hierarchical structure into a flattened representation by capturing higher-order interactions among codes at the same level, enhancing the representation’s richness. For instance, a patent classified with G06F40 and H04W4 is represented as [G, H, G-H, G06, ..., G06F40, H04W4, G06F40-H04W4].

⁵For example, in the case of LLMs, "Transformer" refers to a revolutionary architecture proposed by Vaswani et al. (2017). However, "transformer" can also describe the widely used electrical device for changing voltage when transferring electric energy from one alternating-current circuit to another. A subject matter expert can immediately differentiate the two upon first encounter, whereas a language model requires a substantial amount of data to capture the differences.

⁶The same methodology can be applied to other popular patent classification systems, including IPC and WIPO.

This technique is analogous to incorporating n -grams in the *Bag-of-Words* representation of textual documents (Gentzkow et al., 2019). The subsequent step aggregates the patents at the firm level, creating a firm-token matrix where a firm’s overall R&D is represented by the frequency of each token (e.g., G-H) in its patent (application) portfolio.

The constructed firm-token matrix, being sparse and high-dimensional, is not immediately conducive to depicting firms’ technologies. Subsequently, dimensionality reduction, following row-normalization, compresses this sparse representation into a denser, lower-dimensional matrix ⁷. This process builds on the classic Latent Semantic Analysis technique, introduced in Deerwester et al. (1990).

Algorithm 1 Creating Latent Technology Space

1. Specify D , the desired level of depth in the hierarchical patent classification system.
 2. Specify n , the max order of interaction among classification codes in the same level of hierarchy.
 3. *for* d in $\{1, \dots, D\}$:
 - 3.1 Define the set of tokens by interacting classification codes up to the n -th order;
 - 3.2 Create the patent-token counts matrix;
 - 3.3 Aggregate the counts matrix at the applicant level; store for the next step;
 4. Concatenate and normalize the applicant-count matrices.
 5. Apply dimensionality reduction.
 6. (optional) For a particular technology:
 - 6.1 Find the related patents.
 - 6.2 Create the patent-token matrix of counts.
 - 6.3 Sum across all patents.
 - 6.4 Apply the transformation used in step 5.
-

The next goal is to determine the position of LLM technology within the created latent technology space. For this purpose, all patents citing Vaswani et al. (2017)⁸, which intro-

⁷In this exercise, I utilized the first four levels of the CPC system to represent firm technologies, for instance, up to G06F40 as demonstrated in the previous example. I included only applicants with at least five applications between 2019 and 2023, resulting in nearly 1.25 million patent applications. Furthermore, I discarded code combinations occurring less than five times due to their rarity. After aggregation at the firm level, the resulting matrix contained nearly 24,000 rows (each representing an applicant) and over 260,000 columns, each corresponding to a technology code combination. Ultimately, I applied truncated SVD to reduce the original matrix to 512 dimensions. This reduced matrix accounts for 69.7% of the variance in the original matrix.

⁸This citation data is available only for granted patents.

duced the Transformer architecture, a fundamental building block of LLMs, were collected. I then identified a subset of these patents associated with CPC code G06F40, which denotes handling of natural language data, and treated these as LLM-related patents. These patents were aggregated as though filed by a single hypothetical firm and were projected onto the latent technology space using the previously acquired compression transformation.

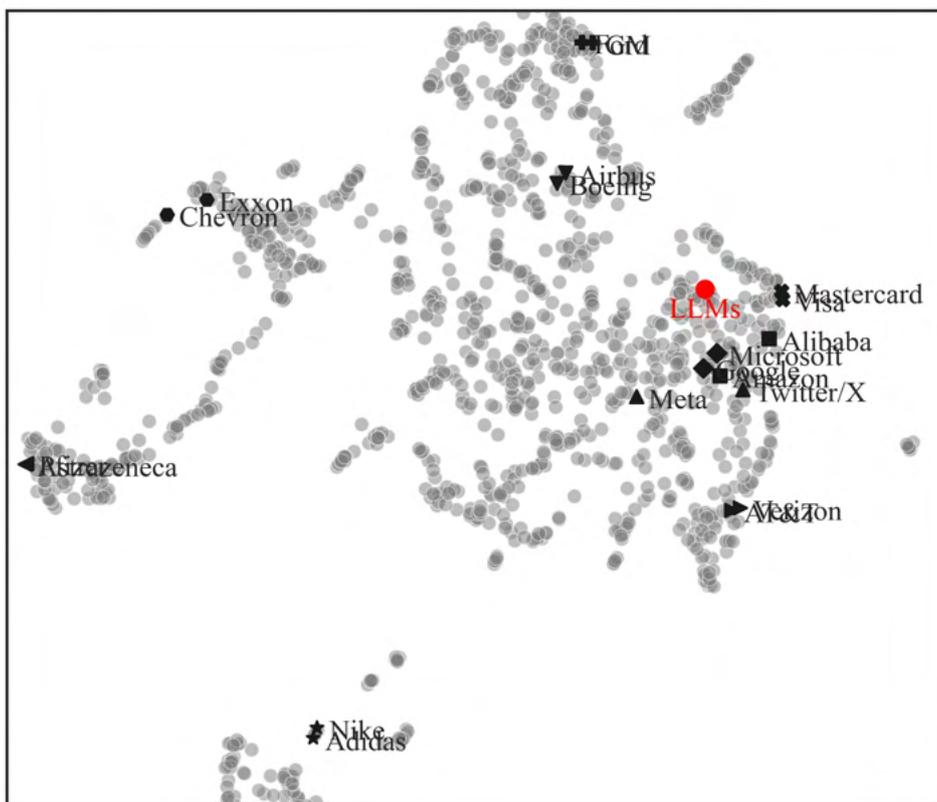
Figure 1.2 illustrates a 2-D representation of the technology space⁹. As a sanity check to verify the proposed method is effective in capturing technology similarities among firms, the figure marks ten well-known pairs of firms with similar R&D portfolios, including Airbus and Boeing, AstraZeneca and Pfizer, and Ford and General Motors. As shown in the figure, these paired firms are positioned in close proximity to one another on the map. Additionally, the figure plots the location of LLM-related patents. As expected, major technology companies such as Amazon, Microsoft, and Google are all positioned close to the LLM technology on the map.

1.4.2 Firms' Compatibility with LLMs and open-source Contributions

Figure 1.3 displays the number of firms in the patent dataset that have an R&D profile compatible with a selected subset of recent technologies, including LLMs. To obtain technology vectors (except for LLMs, whose technology vector was obtained earlier), patents in the dataset corresponding to the CPC code associated with each technology were collected¹⁰. These patents were then processed and aggregated as if filed by a single entity and mapped onto the latent technology vector space, following a procedure similar to that

⁹UMAP package was used to visualize the constructed vector space. To improve the illustration, only firms 100 applications or more are displayed in the figure.

¹⁰The CPC (Cooperative Patent Classification) codes corresponding to the technologies mentioned are as follows: Additive Manufacturing, B33Y10; Computer Vision, G06V10; Cosmonautic Vehicles, B64G1; Cryptocurrency, G06Q2220; Fusion Reactors, G21B; Mixed Reality, G06T19/006; Nanobiotechnology, B82Y5; Quantum Computing, G06N10; Robots, Y10S901.

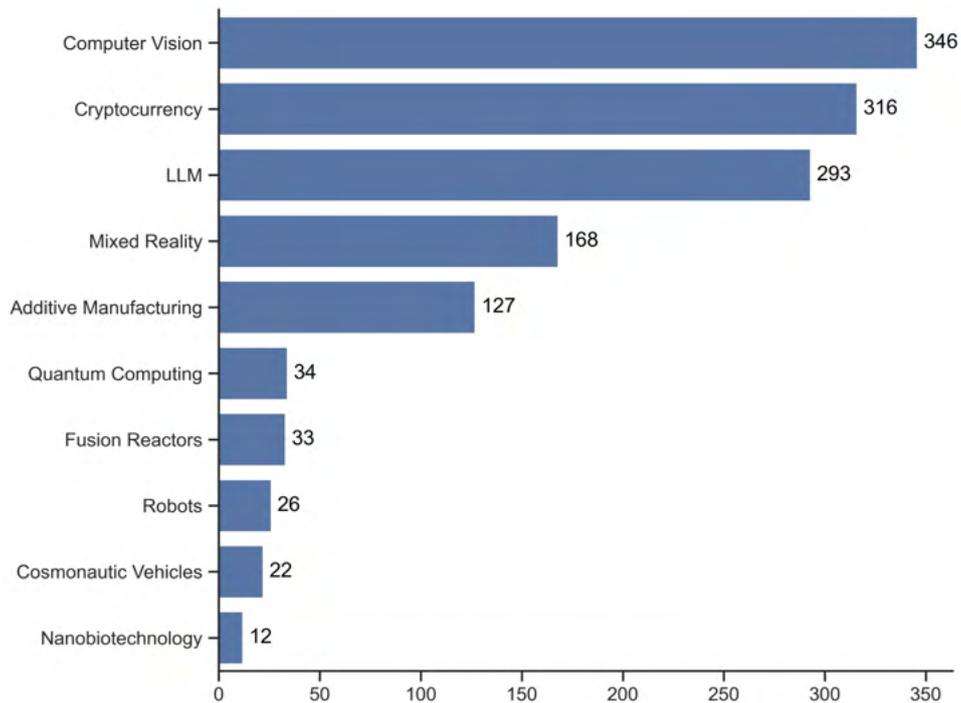


Notes: The figure presents the projection of the constructed latent technology space in 2D. For improved illustration, only applicants with more than 100 applications are included, and a handful of outliers are omitted. Additionally, the figure plots 10 pairs of well-known firms with qualitatively similar technologies. A pooled portfolio of patents citing the “Transformer” paper (Vaswani et al., 2017) and related to natural language processing is represented by a red dot.

Figure 1.2: A 2-D Representation of the Latent Technology Space

described for LLMs. The compatibility of firms with each technology was assessed by cosine similarity between the firm’s vector and the technology vectors, using a critical cosine similarity threshold of 0.7 to distinguish firms with an R&D profile compatible with the technology from those that are not. This process identified LLMs, along with Computer Vision, Cryptocurrency, Mixed Reality, and Additive Manufacturing as technologies compatible with the R&D processes of a relatively large number of firms. However, Quantum Computing, Fusion Reactors, Autonomous Robots, Spacecraft, and Nanobiotechnology were found to be compatible with a smaller subset of firms.

Figure 1.4 illustrates the distribution of cosine similarities between technology vectors

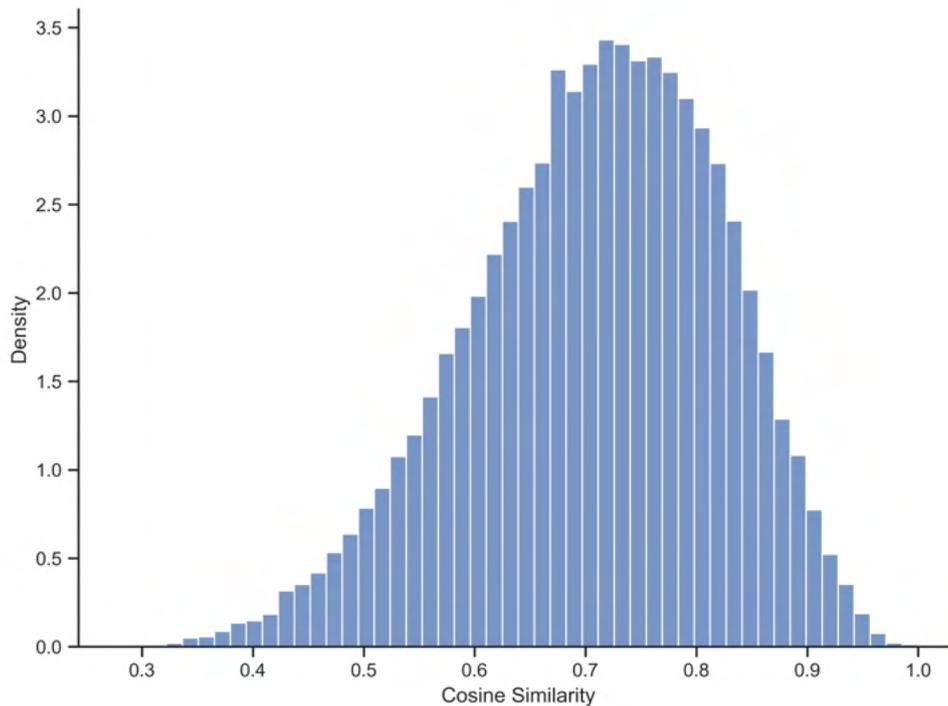


Notes: The figure presents the number of firms in the patent application dataset with R&D profiles compatible to the selected technologies. A firm was considered to be compatible with a technology if the cosine similarity between its R&D vector and the technology vector in the latent technology vector space surpassed a critical threshold of 0.7.

Figure 1.3: Number of Firms with R&D Profiles Compatible to Selected Technologies

of all pairs of firms with LLM-compatible R&D portfolios, as identified in the previous exercise. The figure reveals substantial heterogeneity in the R&D portfolios of firms with LLM-compatible technologies. Additionally, Figure A.1 in the Appendix showcases 50 firms with the largest cosine similarities to LLMs in the latent technology space. Grammarly, an English writing assistance application, has the highest cosine similarity to this technology. The list also includes AI startups and established firms in various sectors, such as Accenture, Baidu, PwC, Thomson Reuters, and Xiaomi. Overall, these findings suggest that LLMs have a broad range of applications in industry, a key assumption for setting up the model in Section 1.6.

Table 1.1 showcases ten companies with the most official repositories of LLM-related papers. The list includes five commonly recognized Big Tech firms: Microsoft, Google,



Notes: The figure presents the distribution of cosine similarities between technology vectors of firms with LLM-compatible R&D profiles.

Figure 1.4: Distribution of Cross Cosine Similarities of Technology Vectors for LLM-Compatible Firms

Meta, Amazon, and Nvidia, as well as other prominent corporations including Alibaba, Salesforce, IBM, Intel, and Tencent. According to the proposed compatibility metric, all these firms have fairly high compatibility with LLMs, and yet none of them are ranked among the top 100 firms with the most LLM-compatible technologies. However, the vast R&D portfolios of these firms, which include thousands of patent applications, imply a significant overall exposure to LLMs. Notably, IBM leads in the number of applications related to natural language generation across the dataset, with Google and Microsoft following in second and fourth places (trailing behind Capital One), and Meta taking the seventh rank. Another notable observation concerns the unique CPC codes within applications related to natural language generation. For instance, Microsoft's 34 applications in this domain encompass 141 unique CPC codes, constituting 12% of all CPC codes in such applications. For IBM, which has the highest number of related applications, this

proportion does not surpass 25%. This observation suggests that even for leading technology firms, the majority of applications related to LLMs may fall outside their R&D scope. Related to this observation, the theoretical analysis suggests that the incentives for open-sourcing advanced software related to a multi-purpose technology are strongest when a firm possesses an intermediate number of compatible applications.

Company	LLM Repos	LLM Compatibility	Patent App.	NLG App.	NLG CPC	Share CPC NLG
Microsoft	199	0.73	7,752	34	141	0.12
Google	118	0.75	7,664	47	155	0.13
Meta	91	0.58	2,834	27	130	0.11
Alibaba	56	0.60	2,242	3	14	0.01
Salesforce	48	0.75	1,834	17	61	0.05
IBM	36	0.76	19,142	115	301	0.25
Amazon	30	0.67	1,840	4	15	0.01
Nvidia	18	0.70	1,851	1	5	0.00
Intel	16	0.52	11,090	4	22	0.02
Tencent	13	0.57	4,486	15	97	0.08

Notes: The table presents selected statistics for 10 firms with the largest number of repositories of LLM-related papers (LLM Repos) on GitHub. ‘Transformer Sim.’ denotes the cosine similarity with patents citing the Transformer paper (Vaswani et al., 2017). ‘Pat. App.’ refers to the number of patent applications filed by that applicant within the dataset. ‘NLG App.’ indicates the number of applications with the CPC code G06F40/56, which denotes natural language generation. ‘NLG CPC’ represents the total number of unique CPC codes co-occurring in natural language generation patents, and ‘Share CPC NLG’ quantifies the firm’s share of all such CPC codes.

Table 1.1: [MODIFY]AI Patent Applications and AI Repositories Owners

1.5 Empirical Analysis

1.5.1 Model Quality and Open-Sourcing Decisions

I start this section by examining how the quality advantage of LLMs over leading open-source alternatives influences the developers’ open-source decisions. The analysis uses models from the Ecosystem dataset, provided by the Center for Research on Foundation Models (CRFM) at Stanford University, that have MMLU scores available. The MMLU

is a widely-used benchmark to assess the general performance of LLMs¹¹. Figure 1.5 illustrates how the leading LLMs' performance on this benchmark has evolved over time.

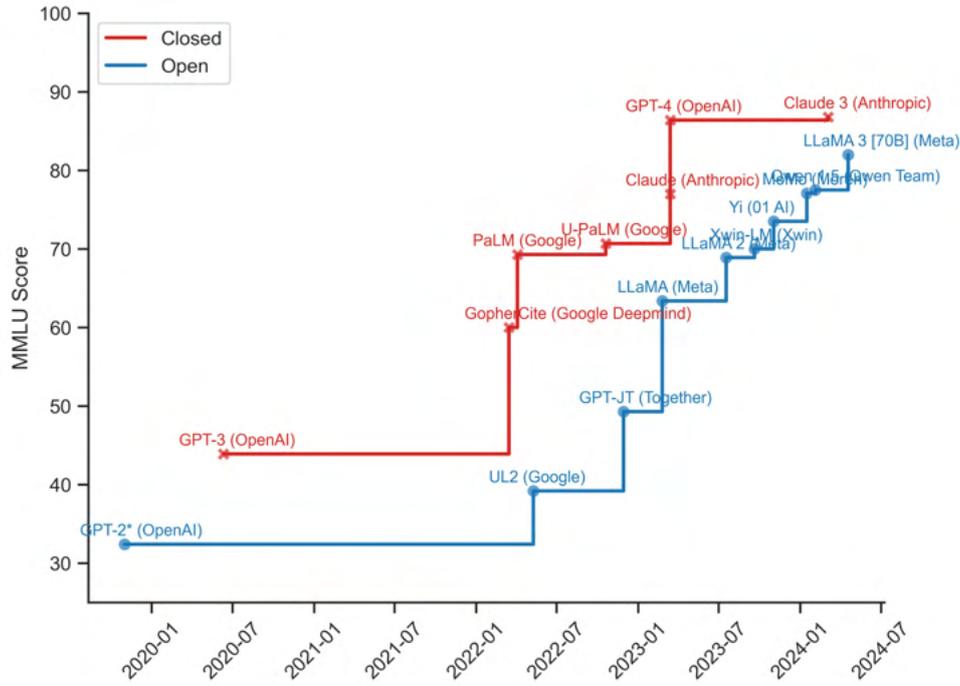
As displayed in Figure 1.5, there has been a persistent gap in performance quality between proprietary and open-source models. Early LLMs, such as OpenAI's GPT-2, were primarily open-sourced and used for research purposes. By contemporary standards, these early models had limited capabilities. GPT-3, a pioneering proprietary LLM, was significantly more advanced than its open-source counterparts at the time. OpenAI's decision to adopt a proprietary release strategy for GPT-3 is aligned with the predictions of the theoretical framework in Section 1.6, suggesting that a LLM developer will adopt a proprietary release strategy if the model's lead over its open-source alternative is large enough. Since the release of GPT-3, closed LLMs have stayed ahead of the curve. Nevertheless, high-quality open-source LLMs have narrowed this gap between open and closed models¹².

Further suggestive evidence worth noting concerns the developers of frontier models. Most top-tier closed LLMs have been released by a few organizations, particularly Google, OpenAI, and Anthropic. Considering the scaling laws of LLMs (Kaplan et al., 2020), a model's performance is primarily determined by its size, training data, and computational resources. Therefore, training LLMs that can outperform previous state-of-the-art models tends to be increasingly costly and out of reach for organizations without substantial resources. Nevertheless, the open-source ecosystem has shown greater dynamism in releasing models that surpass previous frontiers. This greater dynamism can partly be

¹¹The Massive Multitask Language Understanding (MMLU) benchmark comprises a diverse set of natural language understanding tasks, assessing a model's proficiency across various subjects and question types. The random guess baseline score is 25. The scores were sourced from the LMSYS Chatbot Arena Leaderboard (Chiang et al., 2024), Paperswithcode.com, the Huggingace Open LLM Leaderboard, and the models' release reports; they reflect the "5-shot" performance of the models on the benchmark.

¹²For domain specific tasks such as coding or mathematical reasoning, fine-tuned models have already achieved comparable performance to the state-of-the-art closed models. For example, GPT-4 surpasses LLaMA and LLaMA-2 by wide margins on GSM8K (mathematical reasoning) and HumanEval (code generation) benchmarks. However, two LLaMA-based models, MathCoder and WizardCoder, respectively, approach or even slightly exceed GPT-4's performance on GSM8K and HumanEval, according to the models' documentation available at the time of their release.

explained by the fact that, contrary to the closed paradigm, in the open-source ecosystem developers can build on each other’s efforts, leading to more frequent breakthroughs in state-of-the-art models.



Notes: The figure depicts the evolution of the performance of open and closed frontier LLMs in the CRFM data as measured by the Massive Multitask Language Understanding (MMLU) benchmark. A frontier open (closed) model is defined as a model that outperforms its preceding open (closed) models on this specific benchmark. The name of the developers are provided in parentheses. *The scores for GPT-2 reflects the score of the fine-tuned model.

Figure 1.5: Quality Evolution of Frontier Open and Closed LLMs

To further examine the relationship between model quality and open-sourcing decision, consider the following regression,

$$y_i = \alpha + \beta (Q_i - Q_{O,t}^*) + \gamma X_i + \varepsilon_i \quad (1.1)$$

where y_i is a binary outcome equal to one if model i is open-sourced. The main independent variable of the regression is $(Q_i - Q_{O,t}^*)$ that shows the difference between the quality of model i and the quality of the best available open-source model. The theoretical

framework predicts that β is negative. That is, ceteris paribus, if a model surpasses its existing open-source alternative by a wider margin, the owner is less likely to open-source it.

The main challenge for estimating the above regression is that there is no universally available and agreed upon measure of quality for LLMs. Even widely-used benchmarks like MMLU are available for only a subset of models in the CRFM dataset, where score availability is likely influenced by model quality. Nevertheless, if a model's performance is inferior to that of a comparable top-tier open-source model, marginal quality improvements are unlikely to influence the owner's decision to open-source. Hence, my focus is on top-tier models, where benchmark score data are more readily available and the relationship between model quality and open-sourcing decisions is most relevant.

Table 1.2 presents the linear-probability-model (LPM) estimates of the parameter of interest β . As expected, all estimates of β have a negative sign, suggesting that a larger gap between a model and the best available open-source option decreases the likelihood that the model will be open-sourced. Furthermore, the estimates of β suggests economically significant correlations. A 10-point (out of 100) increase in the performance of the model on MMLU with respect to the best available open-source option is associated with a 10-11 percentage point decrease in the likelihood of being open-sourced. The estimates of β change only marginally after including the level of reported (predicted) model quality. The estimates of the level variable are statistically indistinguishable from zero and economically negligible, indicating that the level of model quality is only weakly correlated with the decision to open-source, once the quality difference between the model and the leading open-source alternative is considered.

Unsurprisingly, the coefficients for *For-Profit* organizations' dummy variables in columns (3-4) are negative and statistically significant, indicating that for-profit organizations are,

	(1)	(2)	(3)	(4)
$Q - Q_{o,t}^*$	-0.011*** (0.002)	-0.011*** (0.003)	-0.010*** (0.003)	-0.010*** (0.003)
Q		-0.000 (0.003)	0.000 (0.003)	0.000 (0.003)
<i>For-Profit</i>			-0.138*** (0.0508)	-0.178*** (0.0603)
<i>Big-Tech</i>				0.210** (0.0940)
R^2	0.312	0.312	0.339	0.372
N	86	86	86	86

Notes: The table presents the linear probability model regression estimates of relationship between model quality and open-sourcing decision. Q is the reported (estimated) quality of the model measured by reported (estimated) performance on the MMLU benchmark. $Q_{o,t}^*$ is the quality of the state-of-the-art open-sourced model at the time of model’s release. For the first open-sourced model, the state-of-the-art is considered to be random guess baseline of 25. *For-Profit* is a dummy variable for a for-profit developer. *Big-Tech* is a dummy variable showing if the model is released by one of the following corporations: Google, Meta, and Microsoft. Heteroscedasticity robust standard errors are displayed in parentheses.

Table 1.2: Model Quality Lead and Open Sourcing Decision

on average, less likely to open-source their models¹³. Conversely, the coefficient for the *Big-Tech* dummy variable is positive, large, and statistically significant¹⁴. This finding is aligned with theoretical results, predicting that, ceteris paribus, Big Tech companies are more inclined to open-source their models as they have more compatible applications that can benefit from the positive spillovers of the open-source community.

1.5.2 Open Source as an R&D Catalyst

On February 24, 2023, Meta introduced its large language model, named LLaMA, and made the model available to researchers in academia, industry, government, and civil

¹³The status of each organization was determined manually using online sources such as firms’ websites and Crunchbase. Model size—defined as the number of parameters—is documented for all open models and most closed models.

¹⁴*Big-Tech* indicates if a model is released by Google, Meta or Microsoft. Other recognized Big Tech companies do not have a model included in the dataset.

organizations (Meta, 2023a). This section studies the influence of LLaMA on the activities of LLM researchers. Using activity on GitHub as a proxy for LLM researchers' efforts, I document a significant increase in research-related activities following the release of LLaMA. I must acknowledge the difficulty in making causal claims due to the active period of LLM research around the time of LLaMA's release and the absence of direct data on researchers' use of LLaMA. Nevertheless, the main finding is highly consistent across various specifications, estimators, time horizons, and methods of identifying LLM researchers. It also withstands multiple falsification checks, suggesting a potentially causal interpretation.

A key aspect of open-sourcing LLaMA was Meta's decision to not only provide the fine-tuned assistant model but also the code and parameters of the pre-trained model¹⁵ (see Section 1.2). This enabled researchers to leverage a high-quality pre-trained LLM to tailor it to specific applications, potentially stimulating further research activities around LLMs. Consequently, LLaMA was widely adopted and served as a foundation for a subsequent generation of LLMs¹⁶. However, it remains unclear whether open-sourcing LLaMA merely replaced prior generations of open language models or stimulated further research activity among LLM researchers. This distinction is particularly important as replacing inferior models is equivalent to a one-time upward shift in LLM technology level. However, if open-sourcing LLaMA had a positive influence on R&D activities around LLMs, it could amplify the growth rate of the technology beyond having a positive impact on its level.

Data

Given the notable surge in research on LLMs in 2023, a high-frequency measure of activity is essential to isolate the impact of specific events. Therefore, traditional metrics like the

¹⁵LLaMA-1 was released under a non-commercial license. However, Meta later revised its stance leading to the release of LLaMA-2 in July 2023 under a more permissive commercial license.

¹⁶Meta reported that LLaMA-based models have been downloaded over 30 million times through Hugging Face, and adopted by thousands of startups, innovators, and developers (Meta, 2023b).

numbers of patent applications or publications, which are recorded with significant delays, are not appropriate proxies in this scenario. Consequently, this study utilizes weekly counts of *contributions* on GitHub as an indicator of research activity¹⁷. GitHub defines several activities as contributions, with the primary method being code modifications in a repository (commit). Other forms include code reviews, issue management, and pull requests¹⁸. GitHub also provides information on commits to public repositories, offering a more accurate measure of contributions to open research efforts.

I identified contributors to repositories of LLM-related papers as LLM researchers. To establish a control group, I selected a subset of GitHub users presumably unaffected by the open-sourcing of LLaMA. Given LLMs' significant impact on the broader AI research community, using AI researchers without LLM-related papers for the control group was deemed implausible. Therefore, I identified 20 major repositories on GitHub not directly related to AI, and subsequently collected information of all contributors to those repositories to serve as the control units¹⁹. While it is not possible to verify that open-sourcing LLaMA had no influence on the activities of this group of GitHub users, it is hard to think of a possible scenario in which open-sourcing LLaMA had negative impact on the activity among users in the control group²⁰. As a result, to the extent that open-sourcing LLaMA had a positive effect on the overall activity of users in this group, the results would underestimate the true influence of LLaMA on the activities of LLM researchers.

Similar to other platforms, users on GitHub often include biographical details on their profile pages. These largely unstructured biographies typically feature information about their location, as well as affiliations with universities, companies, or organizations. Given

¹⁷While users can opt to conceal this information, such instances are rare.

¹⁸Pull requests propose incorporating changes from one branch to another, usually for code review and integration before merging.

¹⁹Each chosen repository ranks as the most popular under a specific Topic on GitHub, determined by the number of *Stars*. Repositories solely providing educational materials were omitted. Description of these topics and their corresponding repositories is provided in the Appendix

²⁰The raw data shows a slight increase in the average number contributions among such users in the post period.

the impracticality of manually inspecting the vast number of profiles and the lack of structured data for accurate pattern-based processing, I employed an LLM for data parsing. Specifically, the LLM was tasked with extracting users' countries, their current sector of employment (Academia or Industry), and the names of affiliated organizations, provided this information was available. A manually inspected sample confirmed the LLM's qualitative performance. Further details on utilizing the LLM for data processing can be found in Appendix 1.7. In total, the profiles of over 63 thousand AI researchers were analyzed. The models' prediction suggested that ca. 53% of these researchers work in Academia, 24% work in Industry, and 22% did not disclose this information.

Results

Consider the following event-study regression,

$$y_{i,t} = \alpha_i + \tau_t + \beta_{i,t} Treated_{i,t} + \varepsilon_{i,t} \quad (1.2)$$

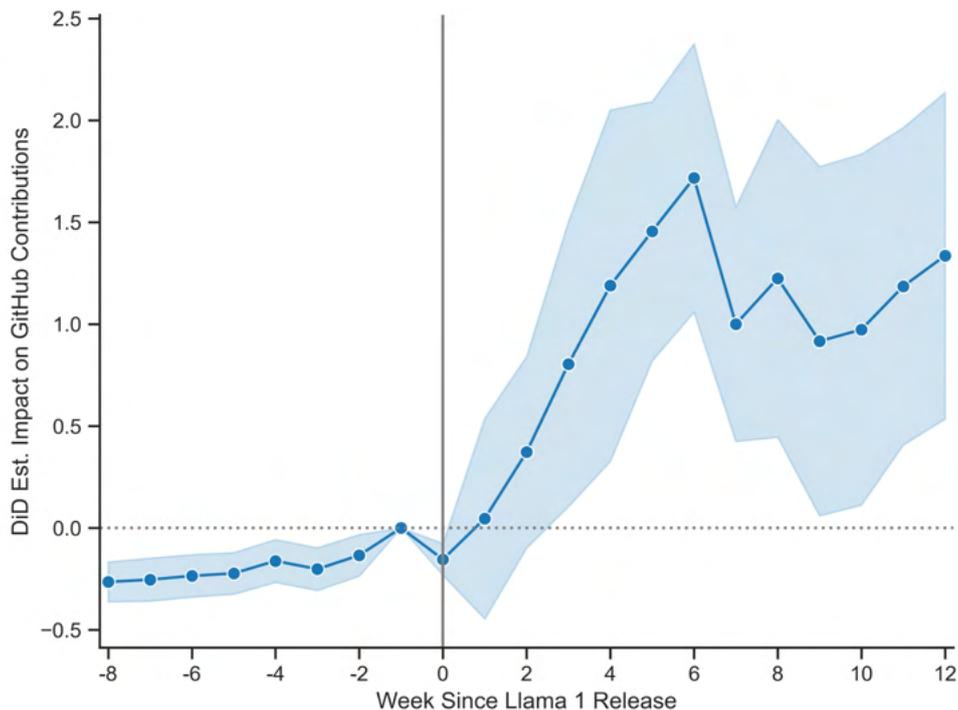
where y represents the relative deviation from the mean pre-event contribution for user i at time t , specifically, $(c_{i,t} - \bar{c}_{i,pre})/\bar{c}_{i,pre}$, where $c_{i,t}$ is user i contributions at time t . This transformation allows interpreting the treatment effect as the average activity level change among affected researchers. Using the raw counts of contributions yields a less intuitive interpretation and skew the results toward the highly active users. Raw contribution counts, while less intuitive and biased towards highly active users, do not alter the robustness of the findings when used as the outcome variable. The analysis includes LLM researchers with at least one pre-period contribution as treated units and non-AI repository contributors as controls. Researchers employed by Meta were excluded. The dataset comprises approximately 6,400 treated and 4,800 control group individuals, respectively.

The goal here is to demonstrate that open-sourcing can stimulate research activities, rather than quantifying the precise impact of LLaMA on the GitHub contributions of LLM

researchers. Contributions on GitHub serve primarily as a proxy for research activity. Therefore, even if a precise treatment effect of LLaMA’s release on GitHub contributions could be estimated, its significance would be limited. Additionally, limitations in the data and identification strategy prevent strong causal claims regarding the treatment effect. Despite these limitations, the findings suggest that open-sourcing an advanced model can do more than replace inferior models; it may catalyze research activity within the community, potentially leading to further advancements and a snowball effect that accelerates technological growth in the field.

Figure 1.6 presents the estimates from the aforementioned event-study regression. The analysis spans a 21-week interval, with Week 0 defined as the seven days following LLaMA’s release on February 24. Notably, the results reveal a moderate pre-trend in the activities of LLM researchers, in comparison to the control group. However, immediately after the model’s release in Week 0, a significant decline in GitHub activities among LLM researchers is observed. This pattern is likely attributable to researchers allocating time to explore the new model rather than contributing to their existing projects. Subsequent to Week 0, LLM researchers’ contributions exhibit an upward trend, stabilizing several weeks later.

Table 1.3 presents the Difference-in-Differences (DiD) estimates of the impact of LLaMA’s open-sourcing on GitHub contributions by LLM researchers (outcomes in Week 0 have been omitted from the analysis). The estimates reveal a substantial and statistically significant increase in the contributions of LLM researchers after LLaMA’s release. This increase is observed for both academia and industry researchers. Including group-specific linear trends has negligible effects on the results, suggesting that the estimates are not primarily driven by pre-trends. The robustness analysis employs the Synthetic DiD estimator proposed by Arkhangelsky et al. (2021) to account for complex pre-trend patterns. The findings are closely aligned with the baseline estimates.



Notes: The figure plots the coefficients from the event-study regression, as described in equation 1.2. The dependent variable is the relative deviation of contributions from their mean pre-event level, defined as $y_{it} = (c_{i,t} - \bar{c}_{i,pre}) / \bar{c}_{i,pre}$. The vertical line marks the introduction date of LLaMA. The shaded areas denote 95 percent confidence intervals, calculated based on standard errors that are clustered at the individual level.

Figure 1.6: Impact of LLaMA on Contributions of LLM Researchers

Furthermore, to ensure robustness, the analysis was narrowed to a 30-day period before and a 17-day period after the model’s release and the results were produced using daily contributions data. This was done to isolate the estimates from the potential effects of GPT-4’s release in March 14. The findings continue to indicate a sizable and significant increase in the contributions of LLM researchers after LLaMA’s release, albeit smaller than the baseline figures. Such a reduced short-term effect was anticipated, as event-study estimates highlighted an increasing momentum in LLM researchers’ activity levels post-LLaMA’s release. Overall, estimates suggest a 40-140% increase in LLM researchers’ contributions after LLaMA’s release, varying with the time horizon and estimation methodology.

Table 1.4 presents the TWFE and Poisson regression estimates of changes in contributions

	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	Academy	Academy	Industry	Industry
ATT	1.202*** (0.196)	1.338*** (0.238)	1.352*** (0.219)	1.453*** (0.269)	0.889** (0.401)	1.127* (0.651)
Obs.	224,240	224,240	166,520	166,520	128,000	128,000
R^2	0.006	0.006	0.007	0.007	0.004	0.004
N. Ind.	11,212	11,212	8,326	8,326	6,400	6,400
Ind. FE	Y	Y	Y	Y	Y	Y
Time FE	Y	N	Y	N	Y	N
Linear Trend	N	Y	N	Y	N	Y
Trend x Treat	N	Y	N	Y	N	Y

Notes: The table presents the Difference-in-Differences estimates of the impact of LLaMA on the activity of LLM researchers on GitHub. The dependent variable is the relative deviation of weekly contributions from their mean pre-event level. The outcomes for Week 0 (the first seven days after LLaMA’s announcement) are omitted. ‘Academy’ indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and ‘Industry’ indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Cluster-robust standard errors are displayed in parentheses.

Table 1.3: Impact of LLaMA on Activity of LLM Researcher on GitHub

by LLM researchers on GitHub, as measured by raw counts of contributions. The results from both estimators are aligned with the baseline results, implying that open-sourcing LLaMA made a positive influence on LLM researchers’ activities on GitHub.

Table 1.5 presents the Difference-in-Differences (DiD) estimates for changes in commit activity to public repositories by LLM researchers. The findings show a significant increase in public contributions on GitHub by LLM researchers following the release of LLaMA. Further analysis indicates that this increase was primarily in the academic sector, with no notable change in public commit activity among industry-based LLM researchers.

Robustness Analysis

Non Linear Pre-Trends

Table 1.3 rules out the possibility that linear pre-trends drive the estimates. Table A.3 presents Synthetic DiD estimates of the treatment effects, offering flexible control over higher-order pre-trends in outcomes between treated and non-treated units. The results

	(1)	(2)	(3)	(4)	(5)	(6)
	All TWFE	All Poisson	Academy TWFE	Academy Poisson	Industry TWFE	Industry Poisson
ATT	0.780*** (0.270)	1.133*** (0.0261)	0.801*** (0.292)	1.168*** (0.0367)	0.795** (0.404)	1.102*** (0.0402)
Obs	224,240	224,080	166,520	166,420	128,000	127,880
R^2	0.004		0.004		0.004	
N. Ind.	11,212	11,204	8,326	8,321	6,400	6,394
Ind. FE	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Trend	N	N	N	N	N	N
Trend x Treat	N	N	N	N	N	N

Notes: The table presents the Difference-in-Differences estimates of the impact of LLaMA on the total weekly contributions of LLM researchers on GitHub. The dependent variable is total weekly contributions. TWFE denotes Two-way fixed-effects estimator, and Poisson denotes the Poisson regression with conditional fixed-effects. The outcomes for Week 0 (the first seven days after LLaMA’s announcement) are omitted. Academy’ indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and Industry’ indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Cluster-robust standard errors are displayed in parentheses.

Table 1.4: Impact of open-source on GitHub Contributions

align closely with the baseline DiD estimates.

Simultaneous Shocks

The first half of 2023 witnessed significant research activity in LLMs. DiD estimates are susceptible to other sources of shocks potentially influencing LLM researchers around the time of LLaMA-1 being open-sourced. Notably, GPT-4, the successor to GPT-3.5 (ChatGPT), was released in mid-March 2023. Despite being an enhanced version of its closed predecessor, GPT-4 introduced additional features, such as multimodality²¹ and advanced code generation capabilities, which could impact research activities. To isolate the results from GPT-4’s potential influence, I use daily data and narrow the time window to 30 days before LLaMA’s release and one day prior to GPT-4’s release. The short-term estimates, presented in Table A.4, remain statistically and economically significant, but are smaller in size compared to the baseline values, likely due to the impact of open-

²¹Multimodality refers to the ability of a single model to process and generate multiple types of data.

	(1) All	(2) Academy	(3) Industry
ATT	0.372*** (0.133)	0.428*** (0.145)	0.274 (0.199)
Obs.	224,240	166,520	128,000
R^2	0.001	0.001	0.001
N. Ind.	11,212	8,326	6,400
Ind. FE	Y	Y	Y
Time FE	Y	Y	Y

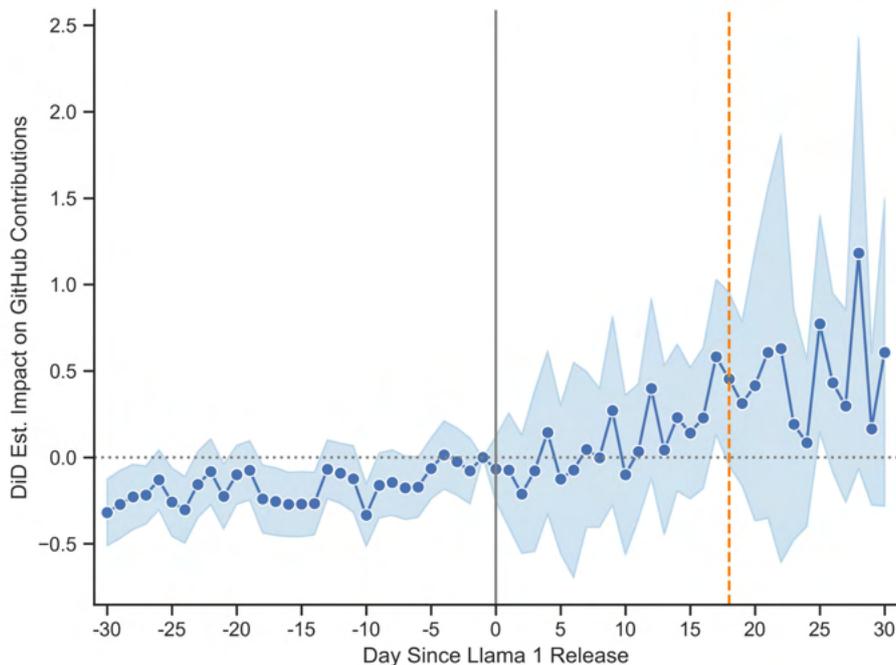
Notes: The table presents the Difference-in-Differences estimates of the impact of LLaMA on weekly public contributions of LLM researchers on GitHub. The outcomes for Week 0 (the first seven days after LLaMA’s announcement) are omitted. ‘Academy’ indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and ‘Industry’ indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Cluster-robust standard errors are displayed in parentheses.

Table 1.5: Impact of open-source on Public Contributions

sourcing LLaMA not being fully realized in the few weeks after its release.

In addition, Figure 1.7 plots the estimated event-study coefficients of Equation 1.2 using daily data around the day of open-sourcing LLaMA. The figure does not reveal any significant increase in the contributions of LLM researchers following the release of GPT-4, and the overall pattern suggests that the trend in contributions remains relatively consistent after open-sourcing LLaMA. Overall, the robustness analysis indicates that the release of GPT-4 is unlikely to account for the main findings.

Submission deadlines of major AI conferences could potentially be another source of shocks. However, such shocks are unlikely to explain the steep rise in the number of contributions after open-sourcing LLaMA and its stabilization after passing 5-6 weeks, as displayed in Figure 1.6. Nevertheless, data on the submission deadlines of two major NLP conferences (ACL and EMNLP) and two major general AI conferences (ICML and



Notes: The figure plots the coefficients from the event-study regression, as described in Equation 1.2. The dependent variable is the daily relative deviation of contributions from their mean pre-event level, defined as $y_{it} = (c_{i,t} - \bar{c}_{i,pre}) / \bar{c}_{i,pre}$. The solid vertical line (in grey) marks the introduction date of LLaMA. The dashed vertical line (in orange) denotes the release of GPT-4. The shaded areas represent 95 percent confidence intervals, based on clustered robust standard errors.

Figure 1.7: Short-Term Impact of LLaMA on Contributions of LLM Researchers

NeurIPS) were collected. The submission deadlines for ACL and ICML fell in late January, before the open-sourcing of LLaMA, while EMNLP’s deadline was in late June, a few weeks after the end of the study period. The submission deadline for NeurIPS on May 17, coinciding with Week 11 in Figure 1.6, appears to be the only potential concern. However, given that the estimate for Week 12 appears to be as large as the estimate for Week 11, it seems unlikely that this deadline had a meaningful impact on the main findings.

Alternative Categorization of the Treated Group

Table A.5 demonstrates the robustness of the results to alternative methodologies for identifying researchers in the treated group. First, papers utilizing the term ‘language model’ in their titles or abstracts are identified, and the contributors to the repositories

associated with these papers are subsequently recognized as the treated group. In the second strategy, a text-based clustering method is employed to discern the largest cluster associated with natural language processing²². Contributors to repositories of papers in the cluster linked to NLP are then determined as the treated group. The results under both methodologies align closely with the baseline estimates.

1.6 Theoretical Analysis

This section introduces a model that explores the dynamics of AI software development and the decision-making process regarding open sourcing. It concentrates on a handful of factors considered to have primary importance, which can be readily incorporated into simple growth models. The goal is to develop a straightforward dynamic discrete choice model to analyze the decisions of a tech firm concerning the open-sourcing a LLM, which serves as an input for producing software applications.

1.6.1 Environment

LLMs as a GPT

The model positions LLMs as an enabling technology that enhances profits in the downstream software application sector (AS). The AS comprises a unit mass of software producers, each differing in LLM-compatibility. Profits from LLM-integration in the AS are contingent on the language model quality, and the amount of *compute*²³ used to integrate the model with that application. Furthermore, using compute improves the model's quality for the subsequent periods.

²²Specifically, K-means clustering is applied to the combined titles and abstracts of papers after dimensionality reduction (utilizing TSVD+UMAP) on the TF-IDF matrix of pre-processed texts. The most frequent cluster among the papers presented at two conferences dedicated to Computational Linguistics (ACL) and Natural Language Processing (EMNLP) is then identified.

²³Compute denotes the total computing resources utilized for algorithm training and execution.

In the AS, software producers are uniformly distributed across the unit interval ($x_i \sim U[0, 1]$), and in each period, they choose to use one type of LLM with quality $q_{\tau,t}$. The profit $\pi_{i,\tau,t}$ for producer i , at time t , spending $k_{i,t}$ units of compute when using model τ with quality $q_{\tau,t}$ is:

$$\pi_{i,t,\tau} = e^{-\gamma x_i} \left(q_{\tau,t} k_{i,t} \right)^\alpha - k_{i,t} - P_{\tau,t} \quad (1.3)$$

where $e^{-\gamma x_i}$ represents the producer's compatibility with LLMs and $P_{\tau,t}$ is the license price of type τ model in period t .

A LLM is either proprietary/closed or open-source. An open-source model is freely accessible to all producers, while proprietary model is priced by its owner at the beginning of each period. Moreover, while only the owner can improve quality of the closed model, the open-source model benefits from collaborative enhancements by all software producers. There are two types τ of models, A and B . Type B is open-sourced with $P_{B,t} = 0$ for all t , while type A is developed and owned by tech firm \mathbf{A} .

The Tech Firm

The model includes a (big) tech firm \mathbf{A} , which owns and controls all software producers in the application sector within the range $x_i \in [0, m]$. While subsequent analysis will explore \mathbf{A} 's decision in developing its own LLM, it is initially assumed that at $t = t_0$, \mathbf{A} is endowed with model A with superior quality compared to the open-source alternative. From $t \geq t_0$ onwards, Firm \mathbf{A} can choose to irreversibly open-source its model, under the condition $q_{A,t} \geq q_{B,t}$ ²⁴.

Open-sourcing allows for external contributions, potentially accelerating the quality growth of model A and, consequently, increasing \mathbf{A} 's profits from internal producers. Alterna-

²⁴This assumption is without loss of generality, as open-sourcing lower-quality model would not influence external producers' choices.

tively, Firm **A** can maintain a closed API, licensing the model to external producers for direct revenue, but this approach excludes the possibility of external quality improvements from the open-source community. The choice between rapid quality enhancement and direct monetization presents a strategic dilemma in open-sourcing decisions.

External Producers

Producers with $x_i \in (m, 1]$ that are not controlled by firm **A** face a static profit-maximizing problem each period. They evaluate the qualities and API prices of all available LLMs types and choose the model that maximizes their profits, based on their optimal compute unit for the selected model type. Specifically they solve,

$$\pi_{i,t,\tau} = \max_{\tau, k_\tau} \left\{ \left(e^{-\gamma x_i} (q_{A,t} k_{i,A,t})^\alpha - k_{i,A,t} - P_{A,t} \right), \left(e^{-\gamma x_i} (q_{B,t} k_{i,B,t})^\alpha - k_{i,B,t} \right) \right\}$$

assuming free access to open-source model B . If Firm **A** chooses to open-source the higher-quality model A , the focus for external producers shifts to optimizing compute levels $k_{A,t}$ for A .

When **A** offers its model through an API, external producers compare the license fee P_t (simplifying the subscript A in $P_{A,t}$) with the quality disparities between A and B in their decision-making process. This results in the following demand relation for the API of model A :

$$Q_{A,t} = \frac{1}{\gamma} \left[\ln \left(\alpha \left(q_{A,t}^{\alpha/(1-\alpha)} - q_{B,t}^{\alpha/(1-\alpha)} \right)^{1-\alpha} \right) - (1-\alpha) \ln \left(\frac{\alpha P_t}{1-\alpha} \right) \right] - m \quad (1.4)$$

The demand relation implies that the demand for model A increases with its quality lead over model B , and decreases with its price P . Note that as Firm **A**'s internal producers have an open access to the model, their mass, denoted by m , is subtracted from the demand relation. For a detailed derivation of the relationship, see Appendix 1.7.

1.6.2 Open Source Decision

Firm **A**'s decision is to either irreversibly open-source model A or maintain its closed status for another period. Consequently, the firm must weigh the value of open model, $V^O(q)$, against the value of closed model, $V^C(q, q_B)$, where q denotes the quality of model A . Notably, the value of open-source model, $V^O(q)$, does not depend on the quality of its open-source alternative, as $q \geq q_B$ implies that every producer would opt for A if it were open-sourced. However, the value of the closed model, V^C , is influenced by q_B , since the demand for A 's API hinges on both the quality of A and its open-source counterpart.

This formulation of the firm's decision-making process aligns well with a dynamic programming model featuring a discrete choice. Consequently, the value of the open model is derived as a solution to the following Bellman equation:

$$\begin{aligned}
 V^O(q) &= \max_{k(x)} \left[\int_0^m \pi(q, k(x)) dx + \beta V^O(q') \right] \\
 \text{s.t. } & q' = q + \psi K_A + \phi K_{-A}
 \end{aligned} \tag{1.5}$$

where β is the time discount factor, ψK_A and ϕK_{-A} reflect the contribution of internal and external computes to the model's quality in the subsequent period, with $\phi, \psi \in (0, 1]$. The integral on the right-hand side of the equation denotes the total profit generated by all software producers owned by Firm **A** in the application sector. Equation 1.5 implies that, in any period, when $k(x)$ is chosen optimally, **A**'s valuation of open model with quality q equals the immediate profit generated by its producers using the model, plus the discounted value of the enhanced model with improved quality q' .

The maximization problem on the right-hand side of Equation 1.5 can be simplified with respect to the optimal compute function $k(x)$. It is important to note that the transition equation depends only on the aggregate internally used compute K_A , and not on the distribution of compute among **A**'s producers. Consequently, when K_A is optimally chosen,

profit maximization implies that marginal profit with respect to k must be equal for any two producers $x_i, x_j \in [0, m]$ owned by \mathbf{A} . After some algebraic manipulation, Firm \mathbf{A} 's aggregate production profit in terms of K_A , can be expressed as:

$$\Pi^F(q, K_A) = \Theta (qK_A)^\alpha - K_A$$

where Θ is a constant term determined by the parameters α , γ and m . For further details on derivations, see Appendix 1.7.

Subsequently, Equation 1.5 can be simplified as

$$\begin{aligned} V^O(q) &= \max_{K_A} [\Pi^F(q, K_A) + \beta V^o(q')] \\ \text{s.t. } & q' = q + \psi K_A + \phi K_{-A} \end{aligned} \tag{1.6}$$

The value of the closed model $V^C(q, q_B)$ depends not only on $\Pi^F(q, K_A)$ but also on the profit from model A 's API, $\Pi^A(q, q_B, P) = PQ_{A,t}$, where $Q_{A,t}$ is the demand for the model's API as given by Equation 1.4. Therefore, the Bellman equation for the closed LLM is²⁵:

$$\begin{aligned} V^C(q, q_B) &= \max_{K_A, P} \left[\Pi^F(q, K_A) + \Pi^A(q, q_B, P) + \beta \max \{V^O(q'), V^C(q', q'_B)\} \right] \\ \text{s.t. } & q' = q + \psi K_A \\ & \& \quad q'_B = q_B + \phi K_B \end{aligned} \tag{1.7}$$

²⁵The equation presents a simplified version of the firm's problem modeled in the numerical analysis. To analyze the model for the complete set of states for q_A and q_B , the complete problem introduces another option which is possibility of switching from own model to alternative open-source LLM after paying some adjustment cost, specifically when $q_A < q_B$. However, as the analysis concentrates on cases where $q_A \geq q_B$, this simplification does not impact the firm's decision regarding open sourcing in the deterministic case. Intuitively, it is not optimal for the firm to forfeit the opportunity to open source its higher-quality model and set the API price so high that it eventually necessitates incurring adjustment costs to switch to an alternative open-source LLM. This intuition is also confirmed by the results of the numerical analysis of the model.

Equation 1.7 implies that if **A** decides to keep its model closed, it gains profits from both production and its model's API, while retaining the option to open-source or keep the model closed in the next period, hence obtaining the discounted value of

$$\max\{V^O(q'), V^C(q', q'_B)\}$$

However, since the model remains closed, its quality in the next period q' only increases due to internally used computes K_A . Additionally, the compute used by external producers using model B enhances the next period's quality of B by ϕK_B . Consequently, when setting the API price, Firm **A** must consider its impact on immediate profits and its future implications via the channel of q_B . An increase in P not only alters Π^A immediately but also affects future profits as producers switching from A 's API to the open source alternative B will contribute to the growth of model B , thereby affecting the demand for A 's API in all subsequent periods.

Finally, Firm **A**'s decision to open-source can be modeled as choosing the option with the highest value:

$$V(q, q_b) = \max\{V^O(q), V^C(q, q_b)\}$$

Analytical Findings

Since the open-sourcing decision involves a discrete choice, an analytical solution does not exist²⁶. However, it is still possible to characterize some key properties of the solution under mild assumptions.

The following proposition establishes that, keeping other variables including q_B constant, if there exists a q^* such that $V^C(q, q_B) = V^O(q)$, then q^* acts as a critical threshold. That

²⁶Dynamic programming models are typically analyzed numerically. Though analytical solutions can be obtained for some special cases using a guess and verify approach, the presence of a discrete choice in the model precludes an analytical solution due to the non-differentiability of the value function.

is, firm **A** will open-source its model only if $q < q^*$ and will keep the model closed if $q > q^*$.

Proposition 1. *Let $q^* > q_{\mathbf{B}}$ be such that $V^{\mathbf{C}}(q^*, q_{\mathbf{B}}) = V^{\mathbf{O}}(q^*)$. Then, the following conditions hold:*

- *If $q > q^*$, then $V^{\mathbf{C}}(q, q_{\mathbf{B}}) > V^{\mathbf{O}}(q)$.*
- *If $q < q^*$, then $V^{\mathbf{C}}(q, q_{\mathbf{B}}) < V^{\mathbf{O}}(q)$.*

Proof. The complete proof can be found in Appendix 1.7. Briefly, the proof involves a first-order approximation of $V^{\mathbf{C}}$ and $V^{\mathbf{O}}$ around q^* , and exploiting that $V^{\mathbf{C}}$ and $V^{\mathbf{O}}$ are strictly increasing with respect to q and $V^{\mathbf{C}}$ is strictly decreasing with respect to $q_{\mathbf{B}}$. \square

The second proposition implies that the firm sets the price of its model's API below its one-period revenue maximizing value to limit the growth of model B .

Proposition 2. *Unless the optimal solution implies $V^{\mathbf{C}}(q', q'_{\mathbf{B}}) < V^{\mathbf{O}}(q')$, the firm sets P below its one-period revenue-maximizing value.*

Proof. The proof involves using the first order condition and the envelop theorem with respect to P and $q_{\mathbf{B}}$. See the complete proof in Appendix 1.7. \square

Numerical Analysis

Dynamic programming models that involve a discrete choice must be analyzed numerically. For this purpose, I use Value Function Iteration (VFI) method, one of the most common methods to solve dynamic programming models in economics. VFI hinges on the assumption that the transformation of the value function is a contraction. This technique begins with an arbitrary initial guess for the value function, which is then iteratively updated using the Bellman equation and guaranteed to converge under a couple of weak

and commonly made assumptions. For additional details on the numerical analysis, please refer to Appendix 1.7.

Table 1.6 presents the baseline choice of parameters used in the numerical analysis. Note that the choices of these parameters are not to be taken literally. The goal here is to see how the tendency to open-source generally varies with changes in a few key factors of interest within a simple framework, but the numerical values are not important.

Description	Notation	Baseline Value
Time Discount Factor	β	0.9
AI Compatibility Parameter	γ	1.0
Shape Production Function	α	0.45
Size Firm A	m	0.2
Efficiency of Internal Development	ψ	0.5
Efficiency of open-source Ecosystem	ϕ	0.5
LLM Development Improvement Factor	λ	5.0
LLM Development Cost Factor	c_D	0.4

Table 1.6: Baseline Choices of Parameters

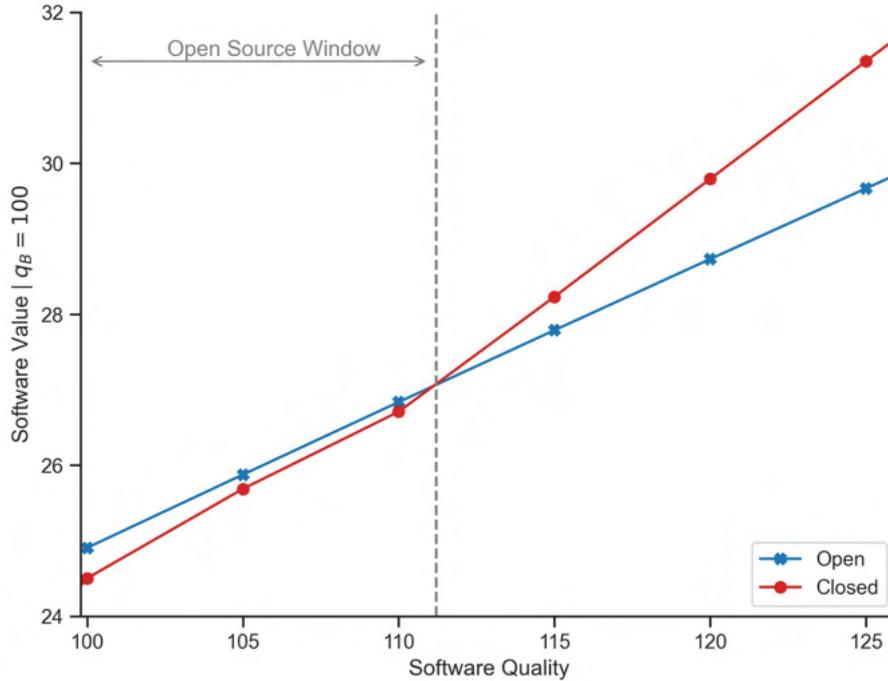
1.6.3 Results

In the following section, I present the results from the numerical analysis of the model.

Open Sourcing Decision

Figure 1.8 illustrates how firm **A**'s valuation of the open and closed model varies with model quality, for a given quality of the alternative model B . Where model A has a modest lead over model B , the value of closed model is smaller than the value of open-source model. In this region, Firm **A** will open source its model. However, there is a critical threshold such that the value of open and closed model intersect. This critical threshold, which is marked by the vertical dashed line in the figure, specifies the open-source window, implying that Firm **A** will open source A only if q_A is in this window, and for the qualities above that critical threshold **A** will keep its model closed. Furthermore,

Figure A.2 demonstrates how the size of the open-source window changes with respect to q_B . The results show that while the absolute size of the window increases for larger values of q_B , its relative size remains roughly constant.



Notes: The figure depicts the valuation of closed and open model for firm **A** as a function of its quality, while maintaining the quality of model **B** at a constant value of 100. The modelling parameters used in the figure are detailed in Table 1.6.

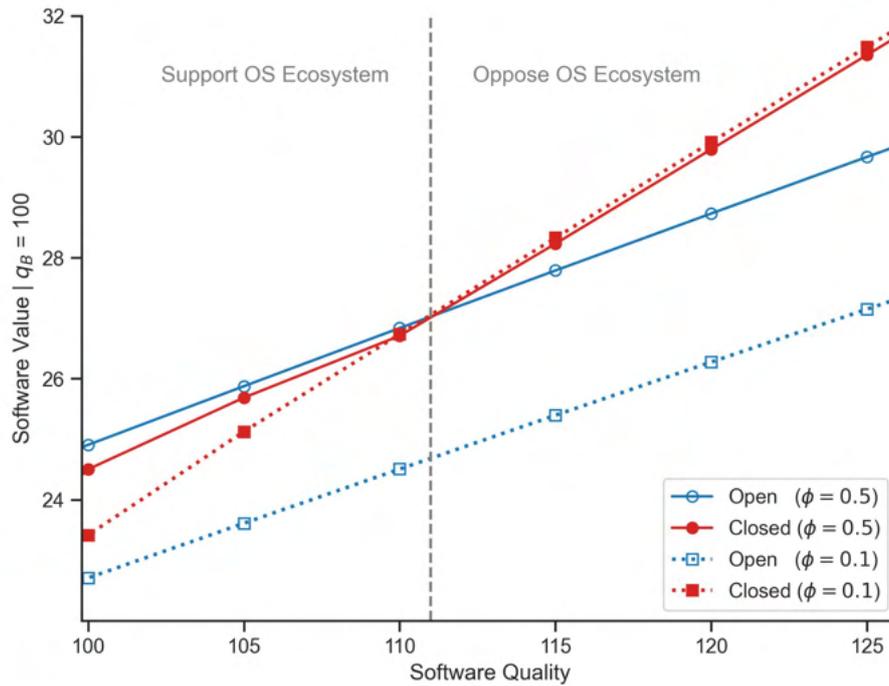
Figure 1.8: Open Sourcing Decision

Efficiency of the Open-Source Community

The efficiency of the open-source community in contributing to model quality, denoted by parameter ϕ , has a straightforward impact on Firm **A**'s open-sourcing *decision*. Essentially, a larger ϕ implies a larger value of the open-sourcing option by accelerating the growth opportunities of open model, and a smaller value of closed-source model by increasing the growth potential of model **B** and decreasing API profits. Therefore, the open-source window size must be increasing in ϕ . Figure A.3 illustrates such a relationship. The figure shows that if ϕ is sufficiently small, then the size of the open-source window is 0, that is, **A** will not open its LLM even if it is only marginally better than **B**.

The more interesting question is how the efficiency of the open-source community affects the overall model *value*, after accounting for its impact on the open-source decision. Figure 1.9 illustrates the impact of a reduction in ϕ , from its baseline value of 0.5 to 0.1, on the value of closed and open model for Firm **A**. As anticipated, the value of open model diminishes with a decrease in ϕ , since in the open-source scenario, the efficiency of the open-source ecosystem accelerates the model’s development, thereby enhancing its value. Conversely, the impact of a reduction in ϕ on the value of closed model is more complex. When the lead of model A over the alternative model B is not too large, a decrease in ϕ also lowers the value of closed model. This decrease stems from the fact that the optimal decision for such q values is to open-source the model, and the value of open model influences the immediate value of closed model through the discounted option value of choosing between open and closed model, represented by $\beta \max\{V^O, V^C\}$ in Equation 1.7. However, if the lead of model A over B is large enough that Firm **A**’s optimal decision is to keep A closed, then a reduction in ϕ essentially increases the value of closed model by limiting the competition between model A and B , thus enabling Firm **A** to extract larger profits by selling API access to its model.

The impact of ϕ on **A**’s model valuation presents another intriguing interpretation. The vertical dashed line in the figure marks the critical q at which the value of the LLM, i.e., $\max\{V^O, V^C\}$, with $\phi = 0.1$ exceeds the model value when $\phi = 0.5$. Consequently, **IF** the firm could influence the efficiency of the open-source ecosystem exogenously, such as by investing in the open-source ecosystem or lobbying for regulations, it would opt for the former strategy when its lead over the open-sourced alternative is small/moderate, and the latter when its lead is substantial.



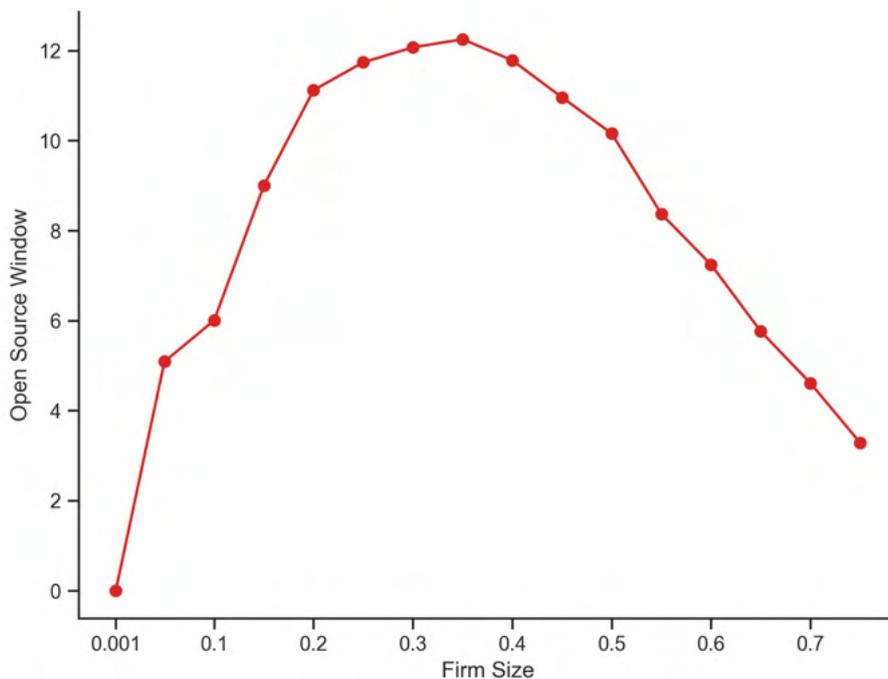
Notes: The figure shows the valuation of closed and open model for firm **A** as a function of its model quality, under two different values of (ϕ) denoting the efficiency of the open-source community. The quality of model *B* is held constant at 100. For further details on the modeling parameters, see Table 1.6.

Figure 1.9: Efficiency in the open-source Community and LLM's Value

Open Source Decision and Firm Size

Subsequently, I explore the interplay between the size of firm **A** and its decision to open-source. Here, "firm size" refers to the mass of producers in the application sector owned by the firm and possessing AI-compatible technology. Figure 1.10 illustrates the size of the open-source window for various firm sizes. As indicated in the figure, the influence of firm size on open-sourcing decisions is not linear. For a relatively small firm, the profits from internal production are minimal, diminishing the advantages of open-sourcing, which primarily lies in the accelerated model quality growth. In such cases, revenue from API licensing becomes significantly more valuable, leading to weak incentives for open-sourcing. Conversely, for a larger firm, internal production profits become a major portion of overall profits, favoring open-sourcing due to the accelerated growth of open LLM and, consequently, the increased profits generated by Firm **A**'s producers in the AS.

However, when the firm becomes too large, the benefits of external contributions to open-source model diminish in comparison to internal contributions, diminishing the incentives for open-sourcing once again. Overall, this results in an inverted-U shaped relationship between the firm's size and its inclination to open-source its advanced LLM.



Notes: The figure illustrates the open-source window size, indicating the maximum quality lead of model *A* over model *B* at which Firm **A** opts to open source model *A*, across different values of firm size m . The quality of model *B* is fixed at 100. See Table 1.6 for additional details on the modeling parameters.

Figure 1.10: Firm Size and Open Source Decision

Development Decision

In the previous analysis, I assumed that the firm is initially endowed with an advanced LLM. This section delves into the decision to develop new model after observing the quality of existing open-source LLM, q_B . Two assumptions underpin the functional form of the new model development process. First, the cost of creating a new LLM is proportional to the current open-source LLM, represented as $c_D q_B$. Secondly, the expected quality of this new model is denoted by $\lambda^u q_B$, where u is drawn from a uniform distribution $U[0, 1]$, and $\lambda > 1$ is the development factor. Thus, the expected value of developing a new LLM

is given by:

$$E(V = \max\{V^O, V^C\}|q_B) = E(V_A(\lambda^u q_B, q_B)) - c_D q_B \quad (1.8)$$

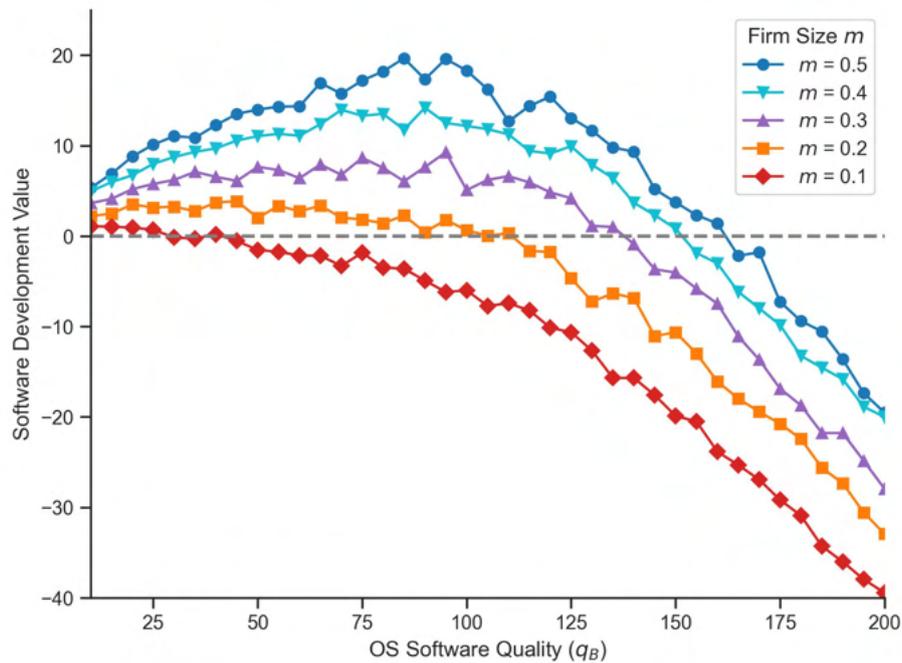
where $E(V = \max\{V^O, V^C\}|q_B)$ indicates that if the firm develops a new model, its value is determined based on the observed quality and its open-source rival's quality, q_B , leading to the selection of either open-sourcing or keeping it closed to achieve the maximum value of the open and closed options.

Figure 1.11 demonstrates how the expected value of developing a new LLM varies with the quality of existing open-source LLM and firm size. The figure reveals that the expected value of developing a new LLM increases with firm size and *generally* diminishes with the quality of the existing open-source model. This pattern suggests a notable dynamic in model development: during the early stages of technology, when existing model quality is low, both small and large firms find it beneficial to invest in developing an advanced alternative. However, as the quality of existing model improves and the investment required for new development escalates, only larger firms continue to find development profitable.

1.7 Conclusion

The primary objective of this study was to explore the rationale behind for-profit companies' decisions to open-source their AI software from a profit-maximizing perspective, focusing on Large Language Models (LLMs) as a particular example.

Analyzing the technology landscape using patent data reveals that LLMs are compatible with the R&D portfolios of a wide array of firms, across varying sizes and with differentiated research technologies. Furthermore, exploiting the open-source release of LLaMA, the LLM developed by Meta, the impact of open-source contributions on stimulating LLM-related research activities was studied. The results suggest that contributions by



Notes: The figure depicts the expected value of developing new AI model relative to the quality of existing open-source model, across various firm sizes. Additional modeling specifics are provided in Table 1.6.

Figure 1.11: LLM Development Decision

LLM researchers on GitHub, considered to be a proxy for research activity, significantly increased following the release of LLaMA.

Additionally, a profit-maximizing firm's decision regarding the development and open-sourcing of a LLM as a multi-purpose technology was modeled in a dynamic discrete choice framework. The theoretical analysis yielded several compelling insights. The predictions suggested that initially, both small and large firms might find it advantageous to invest in developing new LLMs. As the development phase progresses towards the middle and late stages, only larger firms might continue to commit to new model development. Decisions about open-sourcing hinge on the model's quality advantage over open-source rivals and the firm's size. A substantial gap between a firm's model and the previous open-source state-of-the-art acts as a deterrent to open-sourcing, as does being excessively small or large.

Lastly, it's important to acknowledge that this study merely scratches the surface of a complex and evolving topic with significant implications for both industry and policy-making. Future research should delve deeper into areas that remain underexplored in this study. Among these, the influence of competition between AI developers on their open-sourcing decisions present a critical area for exploration. Additionally, the potential impacts of regulating open-source model warrant comprehensive investigation given their far-reaching consequences. Equally important is examining how open-sourcing advancements influence the behavior of downstream firms. These areas represent fertile ground for future studies, promising to enrich our understanding of the ever-changing landscape of AI development and its broader economic and societal impacts.

Appendix

Additional Details on Data

Data Sources, Extraction and Pre-Processing

Papers-with-Code; arXiv

Papers-with-Code is a community-driven initiative providing practitioners with free access to AI/ML research resources. This platform maintains up-to-date information on open-access AI/ML publications and the code repositories associated with those papers. I retrieved the data in January 2024. The raw dataset contains c.a. 219,000 publications (there are instances where a paper has entered the dataset multiple times). The raw dataset does not include the publication dates and the abstracts of papers. Therefore, I use arXiv API to retrieve data of more than 142 thousand publications with unique “arxiv id”. For the main analysis, I use only publications linked with an official code repository on GitHub, and use the publications with unofficial repositories for training the LLM classifier in later stages. I extract the publication date from the ‘published’ field in the arXiv dataset, and keep only papers published from 2019 onward. The following table shows the number of publications in the dataset per year.

Year	N. Papers
2019	11,287
2020	17,734
2021	23,508
2022	26,613
2023	28,755
2024	290

Table A.1: Number of publications in the arXiv dataset

GitHub

In the first step, I retrieve the data for the remaining repositories in the Papers-with-Code dataset from GitHub using its API, of nearly 108 thousand observations, data for c.a. 106 thousand repositories were successfully retrieved. For each of these repositories, I then collect data for owners and contributors to these repositories. Data for more than 109K unique contributors and 8.2K unique organizations was successfully retrieved.

For creating the control group, I also retrieved profiles of data contributors to 20 popular repositories in various topics that are not directly AI-related. Each repository is the most starred non-educational repository associated with a particular “Topic” on GitHub. The following table presents the topic and the link of each repository in the list.

Topic	Repo
Blockchain	ethereum/go-ethereum
Quantum	Qiskit/qiskit
Database	netdata/netdata
Cloud	localstack/localstack
PHP	laravel/laravel
JavaScript	vuejs/vue
Android	flutter/flutter
3D	mrdoob/three.js
IoT	home-assistant/core
Docker	moby/moby
Golang	golang/go
Microservices	nestjs/nest
Django	django/django
Rust	denoland/deno
Game Development	godotengine/godot
Dashboard	grafana/grafana
CLI	ohmyzsh/ohmyzsh
Vim	neovim/neovim
REST	tiangolo/fastapi
Web Applications	angular/angular

Table A.2: Non AI-related Repositories

Patent-Application Data

I used patent application data files with US Patent and Trademark Office. The data were accessed through PatentsView.org in February 2024, containing related data until the end of 2023. The scope was limited to utility patent applications filed by organizations with at least two applications during 2019-2023. The dataset comprises nearly 1.4 million applications. The organizations that their name contained terms indicative of universities or research institutes were removed from the sample. The name of the organizations were cleaned and unique applicant ids were created based of the cleaned names, resulting in more than 60K applicants. The *current* version of Cooperative Patent Classification System was used.

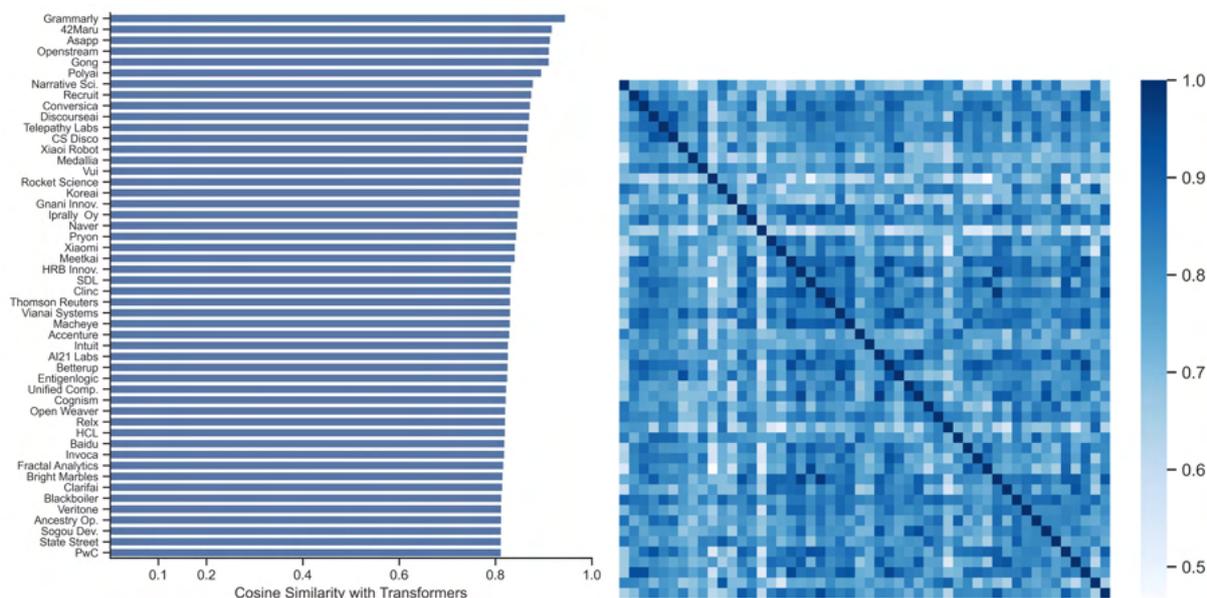
Feature Extraction with LLMs

The scale and unstructured format of data prohibited manual or pattern oriented feature extraction. Therefore, I used LLMs to extract features for three separate tasks. First, given users profile information (bio, location, company) the LLM was asked to extract country, current organization, and sector (academia or industry) if such information is provided in the given text. Second, from the information provided for an organization (name, location, and bio) determine if the organization is a commercial entity. Finally, for creating training data, the LLM was given the title and and the abstract of the paper, and was asked to determine if the paper is related with LLMs at any capacity. For the first two tasks, GPT3.5 API was used to ensure consistency. However, for the final task, to reduce dependency on a single model, data was split between GPT3.5 and Mixtral 7x8B. Each each observation were processed individually, and models' temperature were set to zero to make sure that the results are reproducible.

Additional Results

Firms Compatible with LLMs

Figure A.1, Panel (a), displays the cosine similarity of the 50 companies closest to the Transformer vector in the constructed latent technology space. The firm with the highest similarity to Transformer technology is Grammarly, an English writing assistance application. The list also contains many AI startups as well as established firms such as Xiaomi, Thomson Reuters, Accenture, and PwC. An interesting observation emerges: none of the commonly recognized as “Big Tech” companies are present in the list²⁷. Panel (b) plots the pairwise cosine similarities of these firms’ vectors in the constructed latent technology space. Although these firms are fairly close in terms of their similarity with LLMs, it appears that their overall technology portfolios are quite differentiated.



Notes: Panel (a) presents the 50 firms with the highest cosine similarity to Transformers in the latent technology space. Panel (b) plots the cross-cosine similarity among firms included in the left panel.

(b) Cross-Technology Similarity

²⁷Another interesting observation is that many companies in the list have not filed for either an NLP or a Natural Language Generation application. As indicated by the corresponding CPC codes for NLP, G06F40/4, and NL Generation, G06F40/56.

Synthetic Difference-in-Difference Estimates

	All	Academy	Industry
ATT	1.109*** (0.159)	1.249*** (0.215)	0.769* (0.417)
N	11,212	8,326	6,400

Notes: The table presents the Synthetic Difference-in-Differences estimates of the impact of LLaMA on total weekly contributions of LLM researchers on GitHub. The outcomes for Week 0 (the first seven days after LLaMA’s announcement) are omitted. ‘Academy’ indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and ‘Industry’ indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Bootstrapped cluster-robust standard errors are displayed in parentheses (N=50).

Table A.3: SDiD Estimates of Impact of open-source on GitHub Contributions

Simultaneous Shocks

	(1) All	(2) All	(3) Academy	(4) Academy	(5) Industry	(6) Industry
ATT	0.340*** (0.0992)	0.622*** (0.0847)	0.309** (0.120)	0.593*** (0.109)	0.397** (0.165)	0.753*** (0.174)
Obs.	420,250	420,250	312,543	312,543	244,032	244,032
R^2	0.005	0.002	0.005	0.002	0.006	0.002
N. Ind.	10,250	10,250	7,623	7,623	5,952	5,952
Ind. FE	Y	Y	Y	Y	Y	Y
Time FE	Y	N	Y	N	Y	N
Trend	N	Y	N	Y	N	Y
Trend x Treat	N	Y	N	Y	N	Y

Notes: The table presents the Difference-in-Differences estimates of the impact of LLaMA on the daily contributions of LLM researchers on GitHub, limited to 30 days before and 17 days after the introduction of LLaMA, before the release of GPT-4. The dependent variable is the relative deviation of daily contributions from their mean pre-event level. The outcomes for the first seven days after LLaMA’s announcement are omitted. ‘Academy’ indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and ‘Industry’ indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Cluster-robust standard errors are displayed in parentheses.

Table A.4: Short Term Impact Estimates with Daily Data

Other Treatment Variable Definitions

	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	Academy	Academy	Industry	Industry
	LM	K-Means	LM	K-Means	LM	K-Means
ATT	1.402*** (0.221)	1.241*** (0.246)	1.389*** (0.235)	1.514*** (0.265)	1.231** (0.505)	0.880 (0.558)
Obs.	199,800	199,800	154,920	154,920	120,640	120,640
R^2	0.006	0.006	0.007	0.007	0.004	0.004
N. Ind.	9,990	9,990	7,746	7,746	6,032	6,032
Ind. FE	Y	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y	Y
Trend	N	N	N	N	N	N
Trend x Treat	N	N	N	N	N	N

Notes: The table presents the Difference-in-Differences estimates of the impact of LLaMA on the total weekly contributions of LLM researchers on GitHub. The dependent variable is the relative deviation of weekly contributions from their mean pre-event level. LM' denotes the group of researchers who have used Language Model' in their paper titles or abstracts. K-Means' denotes the group of researchers who have at least one paper in the cluster of NLP papers. The outcomes for Week 0 (the first seven days after LLaMA's announcement) are omitted. Academy' indicates the group of LLM researchers whose GitHub profiles indicate that they are working in academia, and 'Industry' indicates the estimates for LLM researchers whose GitHub profiles indicate they are employed in the industry. Cluster-robust standard errors are displayed in parentheses.

Table A.5: Impact of open-source on GitHub Contributions

Theory Framework Appendix

Model

LLM's Demand Relation

Recall that profit function for producer located at $x \in (m, 1]$ is given by:

$$\pi_{i,t,\tau} = e^{-\gamma x_i} \left(q_{\tau,t} k_{i,t} \right)^\alpha - k_{i,t} - P_{\tau,t}$$

Consider the firm located at point $x \in (m, 1]$ in the AS is indifferent between paying P to access model A 's API and using model B for free. Since by assumption $q_A > q_B$ all producers located in (m, x) will strongly prefer to use model A . Indifference condition for

producer located at x implies:

$$\pi_A = \pi_B \Rightarrow e^{-\gamma x} q_A^\alpha k_A^\alpha - k_A^\alpha - P = e^{-\gamma x} q_B^\alpha k_B^\alpha - k_B^\alpha \quad (9)$$

where k_A and k_B are the optimal compute used when working with model A and B , and given by:

$$k_\tau = \left(\alpha e^{-\gamma x} q_\tau^\alpha \right)^{1/(1-\alpha)}$$

where $\tau \in \{A, B\}$.

From the relations for optimal levels of compute we have $\alpha e^{-\gamma x} (qk)^\alpha = k$. Therefore, we can simplify Equation 9 as:

$$\begin{aligned} \frac{k_A}{\alpha} - k_A - P &= \frac{k_B}{\alpha} - k_B \Rightarrow P = \frac{1-\alpha}{\alpha} (k_A - k_B) \\ \Rightarrow P &= \left(\frac{1-\alpha}{\alpha} \right) \left(\alpha e^{-\gamma x} \right)^{1/(1-\alpha)} \left(q_A^{\alpha/(1-\alpha)} - q_B^{\alpha/(1-\alpha)} \right) \\ \Rightarrow \ln \left(\frac{\alpha P}{1-\alpha} \right) &= \left(\frac{1}{1-\alpha} \right) (\ln \alpha - \gamma x) + \ln \left(q_A^{\alpha/(1-\alpha)} - q_B^{\alpha/(1-\alpha)} \right) \\ \Rightarrow x &= \frac{1}{\gamma} \left[\ln \left(\alpha \left(q_A^{\alpha/(1-\alpha)} - q_B^{\alpha/(1-\alpha)} \right)^{(1-\alpha)} \right) - (1-\alpha) \ln \left(\frac{\alpha P}{1-\alpha} \right) \right] \\ \Rightarrow Q_A &= \frac{1}{\gamma} \left[\ln \left(\alpha \left(q_A^{\alpha/(1-\alpha)} - q_B^{\alpha/(1-\alpha)} \right)^{(1-\alpha)} \right) - (1-\alpha) \ln \left(\frac{\alpha P}{1-\alpha} \right) \right] - m \end{aligned}$$

Aggregate Profit Relation for Firm A's

Since the transition equation in Firm A's dynamic problem only depends on aggregate compute used by all of its producers, the marginal profit of any two producer it owns, w.r.t. compute k , must be equal, otherwise reallocation of one-unit of compute from a producer with lower marginal profit to the one with a higher marginal profit increases

Firm **A**'s profit.

Therefore, consider i and j to be two producers owned by Firm **A** with $x_i, x_j \in [0, m]$.

Following the logic provided above:

$$e^{-\delta x_i} k_i^{\alpha-1} = e^{-\delta x_j} k_j^{\alpha-1}$$

For simplicity assume $j = 0$,

$$j = 0 \Rightarrow k_0 = e^{-\delta x_i/(\alpha-1)} k_i \Rightarrow k_i = e^{-\delta x_i/(1-\alpha)} k_0$$

Therefore aggregate compute can be written as,

$$K = k_0 \int_{x_i=0}^m e^{-\gamma x_i/(1-\alpha)} dx_i = \frac{1-\alpha}{\gamma} k_0 [1 - e^{-\gamma m/(1-\alpha)}]$$

Consequently, we can write aggregate profit of Firm **A** as,

$$\Pi^F(K) = q^\alpha \int_0^m e^{-\gamma x_i} k_i^\alpha dx_i - K$$

Or,

$$\begin{aligned} \Pi^F(K) &= q^\alpha \int_0^m e^{-\gamma x_i} e^{-\delta x_i/(1-\alpha)} k_0^\alpha dx_i - K \\ \Rightarrow \Pi^F(K) &= \left(\frac{\gamma K q}{(1-\alpha)(1 - e^{-\gamma m/(1-\alpha)})} \right)^\alpha \int_0^m e^{-\gamma x_i} dx_i - K \end{aligned}$$

After simplification, we can show that.

$$\Pi^F(K) = \Theta(qK)^\alpha - K$$

Where,

$$\Theta = \left(1 - e^{-\gamma m/(1-\alpha)}\right)^{1-\alpha} \bigg/ \left(\frac{\gamma}{1-\alpha}\right)^{1-\alpha}$$

Proof of Proposition 1

Proof : Suppose for a given q_B and other parameters of the model, there is a $q_A = q^*$ such that the value of open model $V^O(q^*)$ is equal to the value of closed model $V^C(q^*)$. I want to show that for a small Δq , $V^C(q^* + \Delta q) > V^O(q^* + \Delta q)$ if $\Delta q > 0$, and vice versa.

Suppose $\Delta q > 0$, first-order approximation around q^* implies, $V^C(q^* + \Delta q) \approx V^C(q^*) + \Delta q V_q^C(q^*)$, and $V^O(q^* + \Delta q) \approx V^O(q^*) + \Delta q V_q^O(q^*)$. Since by assumption $V^C(q^*) = V^O(q^*)$, $V^C(q^* + \Delta q) > V^O(q^* + \Delta q)$ only if $V_q^C(q^*) > V_q^O(q^*)$.

Now, let's recall the expressions for V^O ,

$$V^O(q^*) = \max_K \left[\Pi^F(q^*, K) + \beta V^O(q^* + \psi K + \phi K_{-A}) \right]$$

Assuming flow of K in any given period is small compared to stock of q ,

$$V^O(q^* + \psi K + \phi K_{-A}) \approx V^O(q^*) + (\psi K + \phi K_{-A}) V_q^O(q^*)$$

Consequently, first-order conditions (FOC) and the Envelop Theorem, imply:

$$\text{FOC 1: } \Pi_k^F(q^*, K^O) + \beta \psi V_q^O(q^*) = 0$$

$$\text{Env 1.: } V_q^O(q^*) = \Pi_q^F(q^*, K^O)$$

Also, for V^C we have,

$$V^C(q, q_B) = \max_{K, P} \left[\Pi^F(q, K) + \Pi^A(q, q_B, P) + \beta \max \left\{ V^O(q^* + \psi K), V^C(q^* + \psi K, q_B + \phi K_B) \right\} \right]$$

After linear approximation and using $V^O(q^*) = V^C(q^*)$,

$$V^C(q, q_B) = \max_{K, P} \left[\Pi^F + \Pi^A + \beta V^C(q^*) + \beta \max \left\{ \psi K V_q^O(q^*), \psi K V_q^C(q^*) + \phi K_B V_{q_B}^C(q^*) \right\} \right]$$

V^C is decreasing with respect to q_B . Therefore, $V_{q_B}^C < 0$.

Now, assume that $V_q^O > V_q^C$. Hence, $\max \left\{ \psi K V_q^O(q^*), \psi K V_q^C(q^*) + \phi K_B V_{q_B}^C(q^*) \right\} = \psi K V_q^O(q^*)$. And we can rewrite $V^C(q, q_B)$ as,

$$V^C(q, q_B) = \max_{K, P} \left[\Pi^F(q, K) + \Pi^A(q, q_B, P) + \beta V^C(q^*) + \beta \psi K V_q^O(q^*) \right]$$

The FOC w.r.t K implies,

$$\text{FOC 2 : } \Pi_K^F(q^*, K^C) + \beta \psi V_q^O(q^*) = 0$$

However, from the FOC of open model we know: $\Pi_k^F(q^*, K^O) + \beta \psi V_q^O(q^*) = 0$. Therefore we must have,

$$\Pi_k^F(q^*, K^O) = \Pi_k^F(q^*, K^C) \Rightarrow K^O = K^C \Rightarrow \Pi^F(q^*, K^O) = \Pi^F(q^*, K^C)$$

From applying Envelop theorem we have,

$$\text{Env 2: } V_q^C(q^*) = \Pi_q^F(q^*, K^C) + \Pi_q^A(q^*, q_B, P)$$

But from Env 1 and $K^O = K^C$,

$$V_q^O(q^*) = \Pi_q^F(q^*, K^O) = \Pi_q^F(q^*, K^C)$$

Therefore, it must be that,

$$V_q^C(q^*) = V_q^O(q^*) + \Pi_q^A(q^*, q_B, P)$$

As profit from API is increasing w.r.t q , we know $\Pi_q^A(q^*, q_B, P) > 0$. However, $\Pi_q^A(q^*, q_B, P) > 0$ contradicts the assumption we made about $V_q^O(q^*) > V_q^C(q^*)$. Therefore, if V^O and V^C intersect at q^* , then $V_q^O(q^*) < V_q^C(q^*)$.

Since V^O and V^C are both increasing functions of q and $V_q^O < V_q^C$ at any point of intersection, then if q^* exists, it must be unique. Moreover, for any $q > q^*$ $V^C(q) > V^O(q)$ and vice versa. \square

Proof of Proposition 2

Let's first recall the Bellman equation for closed model,

$$\begin{aligned} V^C(q, q_B) &= \max_{K_A, P} \left[\Pi^F(q, K_A) + \Pi^A(q, q_B, P) + \beta \max \{V^O(q'), V^C(q', q'_B)\} \right] \\ &\text{s.t. } q' = q + \psi K_A \\ &\& \quad q'_B = q_B + \phi K_B \end{aligned}$$

First, let's consider there is some $\delta > 0$ such that at the optimal solution $V^C(q', q'_B) + \delta > V^O(q')$. That is the optimal solution implies that Firm **A** will keep the model closed in the subsequent period. Then, the FOC w.r.t P and Envelop theorem w.r.t q_B imply,

$$\text{FOC: } \Pi_P^A(q, q_B, P) + \beta V_{q_B}^C(q', q'_B) \frac{\partial K_B}{\partial P} = 0$$

$$\text{Env.: } V_{q_B}^C(q, q_B, P) = \Pi_{q_B}^A(q, q_B, P)$$

Substituting $V_{q_B}^C$ in the FOC from the Envelop theorem results in,

$$\Pi_P^A(q, q_B, P) + \beta \Pi_{q_B}^A(q, q_B, P) \frac{\partial K_B}{\partial P} = 0$$

However, we know that profits from API is decreasing w.r.t. q_B . Hence, $\Pi_{q_B}^A < 0$. Also, an increase in P results in switching from model A to model B and therefore an increase in K_B , i.e., $\frac{\partial K_B}{\partial P} > 0$. Therefore, for the above equality to hold at the optimal solution, we must have that,

$$\Pi_P^A(q, q_B, P) > 0$$

As Π^A is concave w.r.t. P , the derivation above implies that Firm **A** sets the price of its API below the revenue-maximizing value P^* where $\Pi_P^A(q, q_B, P^*) = 0$.

Conversely, let's assume that there is a $\delta' > 0$ such that at the optimal solution $V^O(q') > V^C(q', q'_B) + \delta'$, Then, the FOC w.r.t P implies,

$$\Pi_P^A(q, q_B, P) = 0$$

Therefore, if the optimal choices in that space implies that Firm **A** must open it's model in the subsequent period, the firm will set the price of its API equal to it's revenue maximizing value. \square

Numerical Analysis

In the numerical analysis of the dynamic programming model, I employed Value Function Iteration (VFI) as the primary method for solving the model. The computational work was executed using Python, with the Numba library to optimize performance. The process involved initially solving the model to determine the value of open source model. This solution then served as a foundational input to subsequently solve the model for the value

of closed model. I discretized the state and control variables into intervals of equal length. Additionally, the producer's grid was discretized over the range $[0, 1]$ with 1000 equally spaced points. The parameters used in the numerical analysis of the open model are detailed in Table A.6.

Description	Value
Model quality lower bound	0
Model quality upper bound	500
Model quality grid size	101
Compute lower bound	0
Compute upper bound	20
Compute grid size	100
Size grid producers	1000

Table A.6: VFI Parameters- Opens Model

In the analysis of the closed model, the value of the open model, derived from the preceding analysis, was used in obtaining the results. Moreover, the analysis of the closed model was substantially more demanding from computational aspects with the introduction of an additional state variables (quality model B) and an additional control variables (API price P). As a result, the computational complexity of the model substantially increases which necessitated special considerations, such as employing smaller grids for the control variables.

Furthermore, an additional choice was introduced in the model to facilitate the analysis for the states where quality of model A was less than model B . This choice involved an additional option for Firm of incurring a cost proportional to q_B to transition from using internal model A to model B . However, this option only influenced the solution in scenarios where q_A was substantially smaller than q_B , which was not the focus of the main analysis about the open sourcing decision of the firm which was conditioned on $q_A > q_B$.

To enhance the efficiency of the control grid usage, for any given state, I determined the maximum P that rendered the producer at $x = m$ (the producer with the highest willingness to pay) indifferent between choosing model A and B . The range from 0 to P_m

was then divided into 20 equal segments. Additionally, I adopted an adaptive approach to define the upper bound of the control variable K , based on the model parameters, ensuring that the maximum K for each model configuration remained within the grid's boundaries for K . The modeling parameters and their specific details are outlined in Table A.7.

For additional details on the numerical analysis of the model and insights into the nuances of its implementation, I invite readers to refer to the accompanying code.

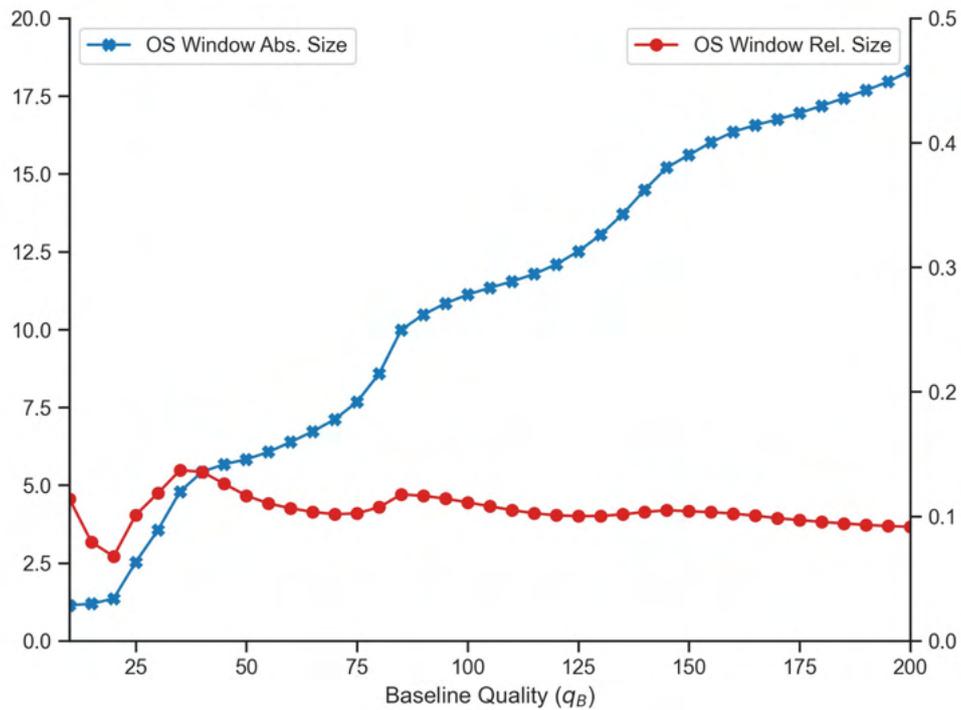
Description	Value
Model quality lower bound	0
Model quality upper bound	500
Model quality grid size	101
Compute lower bound	0
Compute upper bound	<i>adaptive</i>
Compute grid size	50
API price grid size	20
Size grid producers	1000

Table A.7: VFI Parameters- Opens Model

Additional Results

Open Source Window Size and Quality Model B

Figure A.2 shows how the open source window size changes due to changes in the quality of the alternative model q_B . As it is displayed in the figure, the absolute size of the open source window is increasing with q_B . However, the relative size of the window w.r.t q_B is fairly constant.

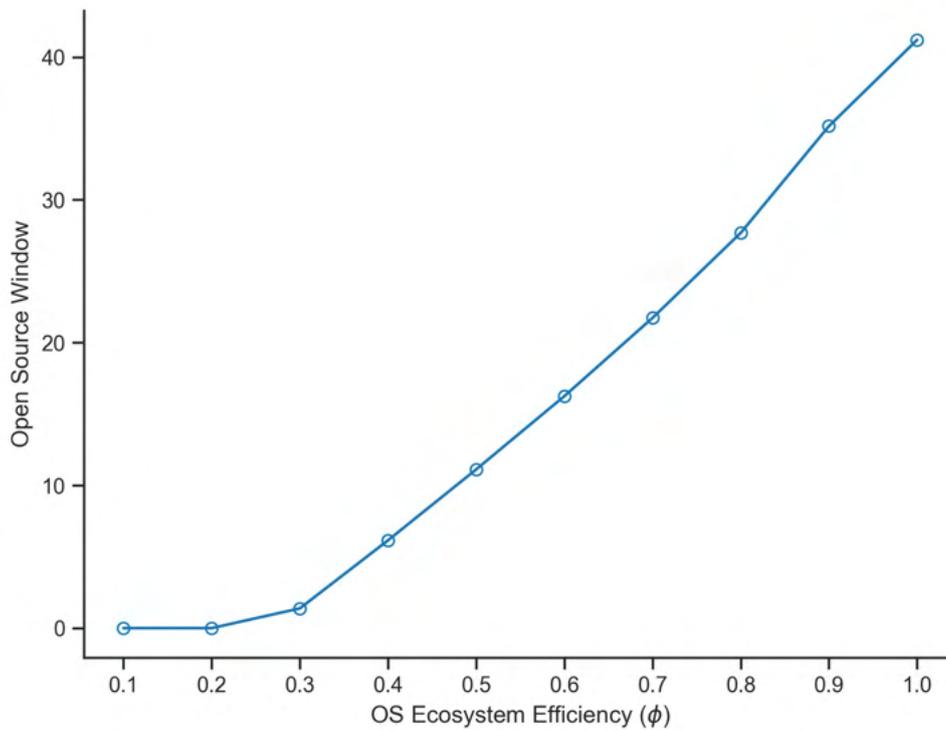


Notes: The figure plots the absolute and relative size of the open source window by quality of the open-source model alternative. The modelling parameters used in the figure are detailed in Table 1.6.

Figure A.2: Quality B and open-source Window A

Open Source Window Size and the Efficiency of open-source Ecosystem

Figure ?? shows how the open source window size changes due to changes in the efficiency parameter ϕ of the open source ecosystem. As expected, the open source window is increasing in ϕ .



Notes: The figure plots the open source window size as a function of efficiency of the open-source ecosystem. The modelling parameters used in the figure are detailed in Table 1.6.

Figure A.3: Efficiency open-source Community ϕ and open-source Window

Bibliography

- Agrawal, A., J. S. Gans, and A. Goldfarb (2023a). Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*.
- Agrawal, A. K., J. S. Gans, and A. Goldfarb (2023b). Similarities and differences in the adoption of general purpose technologies. Technical report, National Bureau of Economic Research.
- Ahmed, N., M. Wahed, and N. C. Thompson (2023). The growing influence of industry in ai research. *Science* 379(6635), 884–886.
- Allen, R. C. (1983). Collective invention. *Journal of economic behavior & organization* 4(1), 1–24.
- Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2021). Synthetic difference-in-differences. *American Economic Review* 111(12), 4088–4118.
- Arora, A., S. Belenzon, A. Patacconi, and J. Suh (2020). The changing structure of american innovation: Some cautionary remarks for economic growth. *Innovation Policy and the Economy* 20, 39–93.
- Arora, A., S. Belenzon, and L. Sheer (2021). Knowledge spillovers and corporate investment in scientific research. *American Economic Review* 111(3), 871–898.
- Arts, S., B. Cassiman, and J. Hou (2021). Technology differentiation and firm performance. *Harvard Business School Strategy Unit Working Paper* (22-040).
- Beltagy, I., K. Lo, and A. Cohan (2019). Scibert: A pretrained language model for scientific text.
- Brynjolfsson, E., D. Rock, and C. Syverson (2018). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In *The economics of artificial intelligence: An agenda*, pp. 23–57. University of Chicago Press.
- Business-Insider (2023). Big tech is inflating fears about ai’s risk to

- humanity: Google brain cofounder. <https://www.businessinsider.com/andrew-ng-google-brain-big-tech-ai-risks-2023-10>.
- Casadesus-Masanell, R. and P. Ghemawat (2006). Dynamic mixed duopoly: A model motivated by linux vs. windows. *Management Science* 52(7), 1072–1084.
- Chiang, W.-L., L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica (2024). Chatbot arena: An open platform for evaluating llms by human preference.
- CNBC (2023a). Meta ceo mark zuckerberg touts to employees ‘incredible breakthroughs’ the company has seen in a.i. <https://www.cnbc.com/2023/06/08/meta-ceo-mark-zuckerberg-talks-companys-ai-efforts-to-employees.html>.
- CNBC (2023b). Meta’s open source approach to ai puzzles wall street, techies love it. <https://www.cnbc.com/2023/10/16/metas-open-source-approach-to-ai-puzzles-wall-street-techies-love-it.html>.
- Cockburn, I. M., R. Henderson, and S. Stern (2018). The impact of artificial intelligence on innovation: An exploratory analysis. In *The economics of artificial intelligence: An agenda*, pp. 115–146. University of Chicago Press.
- Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science* 41(6), 391–407.
- Economides, N. and E. Katsamakas (2006). Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry. *Management science* 52(7), 1057–1071.
- Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2023). Gpts are gpts: An early look at the labor market impact potential of large language models.
- Fosfuri, A., M. S. Giarratana, and A. Luzzi (2008). The penguin has entered the building: The commercialization of open source software products. *Organization science* 19(2),

292–305.

- Gambardella, A. and E. A. von Hippel (2018). Open source hardware as a profit-maximizing strategy of downstream firms.
- Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature* 57(3), 535–574.
- Goldfarb, A., B. Taska, and F. Teodoridis (2023). Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. *Research Policy* 52(1), 104653.
- Hain, D. S., R. Jurowetzki, T. Buchmann, and P. Wolf (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change* 177, 121559.
- Henkel, J. (2004). Open source software from commercial firms—tools, complements, and collective invention. *Zeitschrift für Betriebswirtschaft* 4, 1–23.
- Hovy, D. (2022). *Text analysis in python for social scientists: Prediction and classification*. Cambridge University Press.
- Howard, J. and S. Ruder (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Jacobides, M. G., S. Brusoni, and F. Candelon (2021). The evolutionary dynamics of the artificial intelligence ecosystem. *Strategy Science* 6(4), 412–435.
- Jaffe, A. B. (1986). Technological opportunity and spillovers of r&d: Evidence from firms' patents, profits, and market value. *The American Economic Review* 76(5), 984–1001.
- Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring technological innovation over the long run. *American Economic Review: Insights* 3(3), 303–320.

- Lerner, J., P. A. Pathak, and J. Tirole (2006). The dynamics of open-source contributors. *American Economic Review* 96(2), 114–118.
- Lerner, J. and J. Tirole (2002). Some simple economics of open source. *The journal of industrial economics* 50(2), 197–234.
- Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Meta (2023a, February). Introducing llama: A foundational, 65-billion-parameter large language model. <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>.
- Meta (2023b). The llama ecosystem: Past, present, and future. <https://ai.meta.com/blog/llama-2-updates-connect-2023/>. Accessed: [insert date you accessed the site].
- Nagle, F. (2018). Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. *Organization Science* 29(4), 569–587.
- Nagle, F. (2019). Open source software and firm productivity. *Management Science* 65(3), 1191–1215.
- Nuvolari, A. (2004). Collective invention during the british industrial revolution: the case of the cornish pumping engine. *Cambridge Journal of Economics* 28(3), 347–363.
- NYT (2023, April). Let us show you how gpt works — using jane austen. <https://www.nytimes.com/2023/04/27/upshot/gpt-from-scratch.html>.
- Osterloh, M. and S. Rota (2007). Open source software development—just another case of collective invention? *Research Policy* 36(2), 157–171.
- Post, T. (2023, November). Big tech wants ai regulation. the rest of silicon valley is skeptical. <https://www.washingtonpost.com/technology/2023/11/09/ai-regulation-silicon-valley-skeptics/>.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1(8), 9.

- Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21(140), 1–67.
- Rock, D. (2019). Engineering value: The returns to technological talent and investments in artificial intelligence. *Available at SSRN 3427412*.
- Spencer, J. W. (2003). Firms’ knowledge-sharing strategies in the global innovation system: empirical evidence from the flat panel display industry. *Strategic management journal* 24(3), 217–233.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, Volume 30.
- von Hippel, E. and G. von Krogh (2003). Open source software and the “private-collective” innovation model: Issues for organization science. *Organization science* 14(2), 209–223.
- WSJ (2024). Should ai be open-source? behind the tweetstorm over its dangers. <https://www.wsj.com/articles/should-ai-be-open-source-behind-the-tweetstorm-over-its-dangers-65aa5c97>.

Chapter 2

Content Moderator Dilemma

Mahyar Habibi, Dirk Hovy, and Carlo Schwarz

2.1 Introduction

The widespread proliferation of hateful and inflammatory content online has become an increasing concern for users, policymakers, and online platforms. As growing evidence shows that hateful online content can lead to real-life violence (Müller and Schwarz, 2021, 2022b; Bursztyn et al., 2019; Du, 2023; Cao et al., 2023), platforms have increasingly resorted to content moderation efforts to stem the tide of hateful content online. Prominent examples include the removal of Facebook accounts associated with the far-right group Proud Boys in October 2018 (e.g., NBC-News, 2018), the deletion of Alex Jones’ Twitter account in the aftermath of the Sandy Hook shooting (e.g., BBC, 2018), or most prominently, the suspension of Donald Trump’s Twitter account after the attack on the US capitol on January 6th, 2021 (e.g., Twitter, 2021; NYT, 2021).¹ Also, lawmakers started to introduce regulation of online platforms that demand the removal of Toxic online content. For example, Germany’s “Netzwerkdurchsetzungsgesetz” (BBC, 2017), the

¹Trump’s account was only reinstated after the takeover of Twitter by Elon Musk (The Guardian, 2022) As part of the staff cuts at Twitter, Elon Musk also fired most of the content moderators on Twitter (e.g., USA-Today, 2022)

UK’s “Online Safety Bill”(e.g, Reuters, 2023b), and the “Digital Services Act” of the EU (e.g., Reuters, 2023a) mandate that online platforms are responsible for the content that is circulating on them and therefore have to take content moderation measures.

On the one hand, the concerns about hateful content and the increased demand for content moderation have motivated extensive research on automated hate speech detection (e.g., Davidson et al., 2017; Hanu and Unitary team, 2020; Hartvigsen et al., 2022; Bianchi et al., 2022) and the effectiveness of content moderation efforts (e.g., Chandrasekharan et al., 2017; Jhaver et al., 2021; Jiménez Durán, 2022; Beknazar-Yuzbashev et al., 2022; Jiménez Durán et al., 2022; Müller and Schwarz, 2022a). On the other hand, the expansions of content moderation have been criticized as restrictions to free speech and the plurality of online discourse (e.g., Tworek, 2021; Eidelman and Ruane, 2021; United Nations Human Rights, 2018). In particular, potentially biased applications of content rules have attracted growing criticism from politicians (e.g., Samples, 2019; Vogels et al., 2020; The Texas Tribune, 2022). As a result, online platforms face a dilemma of seemingly contradictory objectives, which they need to balance in their content moderation efforts. Balancing the removal of inflammatory content and the preservation of content plurality is especially difficult since extensive disagreements exist about how these two objectives should be weighted.

It is worth highlighting that this trade-off would *not* disappear even if the “ground truth” of hate speech would be perfectly known, i.e., there would be an unbiased and universally accepted measure of hate speech. Even in this hypothetical scenario, content moderation could still reduce the plurality of online content if specific topics and issues are only discussed using toxic language. The trade-off only vanishes in the unlikely case in which the toxicity of online discourse is entirely unrelated to its content.

This content moderation dilemma is further complicated on the measurement side. While

many methods exist to identify hateful content (see examples above), to the best of our knowledge, no measures exist to quantify the content moderation-induced changes in online plurality. In this chapter, we aim to narrow this gap by putting forward and validating a family of measures for the plurality of online content. We formalize the idea of content plurality as the variance of the semantic space, based on the intuition that the variance relates to the plurality of views.

The measures approximate the semantic space using text embeddings based on Transformer models (Vaswani et al., 2017). The embeddings project texts as representations into a high-dimensional Euclidean space, where their semantic similarity determines their position in relation to all other texts. Texts with similar meanings end up with embeddings that are closer together than unrelated texts. Embeddings have become the de-facto standard in natural language processing (NLP). As of 2024, the original Vaswani et al. (2017) has received over 110,000 citations, and Transformer models have proven highly successful in countless applications (e.g., Zhu et al., 2020; Strudel et al., 2021; Han et al., 2021; Radford et al., 2023), as they have been shown to capture text semantics better than previous approaches. Text embeddings (as opposed to count-based methods) also have proven highly successful in computational social science (e.g., Ash and Hansen, 2023; Garg et al., 2018; Kozłowski et al., 2019).

Our proposed measures of content plurality quantify the aggregate variation of the text embeddings based on metrics of statistics dispersion. Our baseline measure makes use of a Generalized Variance Index (Wilks, 1932) of the embeddings space. Intuitively, the measures increase if the embeddings are more dispersed and decrease if online content becomes more similar. Another advantage of this measure is that it is content-agnostic, i.e., it does not involve any choice of which content is worth preserving. Further, the measure is scalable to large datasets and entire platform ecosystems.

We validate and demonstrate the measure’s potential based on a representative sample of 10 Million US Tweets. Our analysis proceeds in four steps. First, we show that removing toxic online content leads to reductions in the plurality of online content. Importantly, no such reduction occurs if Tweets are removed at random. This result persists independent of the embedding model, toxicity score, or variance measure. To the best of our knowledge, our chapter is the first to provide a measure to quantify this trade-off empirically.

Second, we show that the content moderation-induced reduction in plurality is not driven by the toxic language of the Tweets itself. To this end, we build on the literature on the debiasing of text embeddings (e.g., Bolukbasi et al., 2016), and create projections of the embeddings space that are orthogonal to the toxicity scores. In other words, these projects will assign the same embedding to a text independent of its toxicity. We find that the results remain virtually identical with these modified embeddings. This finding suggests that it is not the toxicity language itself that is important for the variance of the text embeddings.

As the third step, we provide direct evidence for this hypothesis and compare the removal of toxic Tweets with and without content. We use the large language model GPT3.5 to classify whether a Tweet only contains toxic content (e.g., a threat) or contains a message or information (e.g., a view on immigration). This allows us to show that the content moderation-induced reduction in plurality is nearly exclusively driven by the removal of Tweets with content. Removing Tweets containing only toxic language leaves our measure unchanged. This result indicates that our measure of content plurality accurately captures the loss in online content.

Lastly, we propose an alternative version of content moderation that, instead of removing toxic content, rephrases Tweets in a way that strips them of their toxic language while leaving the message intact. For this analysis, we use the language generation capabilities of

large language models. Based on the rephrased Tweets, we can show that this alternative form of content moderation can reduce toxicity while leaving the plurality of online content unaffected.

Our results break new ground by proposing a measure of content plurality that allows us to quantify the trade-off between content moderation and content plurality. Given the fundamental importance of this trade-off, such a measure is of immediate policy relevance. The economic theory of multitask models (Holmstrom and Milgrom, 1991; Feltham and Xie, 1994) predicts that if principles have to choose between different objectives, only one of which is measurable, incentives are such that effort will be focused on the measurable tasks. In other words, in the absence of readily available measures of content plurality, online platforms and lawmakers are likely to put far greater emphasis on removing hateful content at the “costs” of plurality. Hence, our measure represents an essential puzzle piece in the debate on content moderation.

2.2 Data and Methods

2.2.1 Representative US Twitter Data

We use the Tweets from a representative sample of US Twitter users created by (Siegel et al., 2021) for our main analysis. The sample was created by sending queries for user accounts to the Twitter API for random numbers between 1 and 2^{32} , the largest possible Twitter user ID at the time of collection.² If the API returned a user account associated with the random number, the authors confirmed if the user was located in the United States. We collected the Tweets of 432,882 out of 498,901 users whose accounts were still active at the beginning of 2022. In total, this yields a dataset of ca. 400 Million Tweets. We removed non-English tweets, including those containing only links, for our analysis.

²At a later point, Twitter changed the user IDs to 64-bit.

We then draw a 5% random sample from the remaining 220 Million Tweets, leaving us with a final sample of around 10 Million Tweets.³ This final sample should provide a good approximation of the content circulating on Twitter in the United States.

For our content moderation analysis, we create toxicity measures for each Tweet based on two different models. First, our baseline model is Google’s Perspectives API. The Perspectives API has become one of the standard tools for toxicity analysis. For example, it is used by several platforms for content moderation (e.g., NYT, 2016; Forbes, 2019; Delgado, 2019). For each Tweet, the API returns six scores for different toxicity dimensions (toxicity, severe toxicity, identity attack, insult, profanity, and threat). The toxicity scores range from 0 (non-toxic) to 1 (highly toxic) and are roughly interpretable as the share of users who would judge a message as toxic. As is standard in the literature, we will focus on Toxicity scores.⁴ For robustness, we create similar toxicity scores using the Detoxify package (Hanu and Unitary team, 2020). As shown in Appendix A.1 these scores are strongly positively correlated. Additional details on the data and the toxicity measures can be found in 2.4. Appendix A.2 provides examples of toxic Tweets in our data (Warning: The examples contain offensive language.)

2.2.2 Measuring Online Plurality

As described in the introduction, our measure of content plurality is built around the idea of the variance of the semantic space. To build intuition, imagine that the semantic content of a text can be represented as a point in potentially infinite-dimensional semantic space. In this space, texts that talk about the same issue or hold the same opinion are close together, while texts about other issues or diverging opinions are far apart. The diversity (plurality) of content can then be approximated by the variance between the

³As we show in our analysis, the removal of random Tweets has no impact on our plurality measure. The results would be identical if we used all Tweets instead of the 5% subsample

⁴As we show in robustness tests, the results are identical if we use the other toxicity dimensions.

individual texts. The variance will be small if all texts are very similar and, therefore, close to each other in the semantic space. In contrast, if texts are widely dispersed in the semantic space, the variance will be large.

Building on this intuition, we construct our measure of content plurality by first creating an approximation of the semantic space using text embeddings and then calculating measures of variance. We describe each of these steps in the following. First, we create embeddings for each of the Tweets in our data using the BERTweet model (Nguyen et al., 2020).⁵ The BERTweet model, trained on a large English-language Twitter corpus, transforms the text of each Tweet into a 768-dimensional vector. In this way, the BERTweet model generates an approximation of the semantic space.⁶

These types of embedding vectors are crucial for countless NLP tasks, such as text classification, similarity calculation, summarization, translation, generation, and question-answering (e.g., Devlin et al., 2018; Radford et al., 2019; Lewis et al., 2019). In line with the above intuition, the individual dimensions of the embedding vector capture semantic differences between Tweets, i.e., those closer in the embedding space are more similar. At the end of this step, we are left with a $N \times D$ matrix \mathbf{X} , where N is the number of Tweets, and D is the number of embedding dimensions.

As the second step, we use this embeddings matrix \mathbf{X} to construct our measure of content plurality. Our baseline measure of content plurality is defined as the generalized variance index (GVI) (Wilks, 1932) of the matrix \mathbf{X} .⁷ The GVI is a multivariate extension of the standard statistical variance measure based on the mean squared deviation. The GVI is

⁵We also provide robustness checks for embeddings created by the RoBerta and DeBerta models.

⁶We provide additional details on the BERTweet model and embeddings in 2.4

⁷We consider alternative dispersion metrics in robustness checks.

defined as:

$$\text{GVI} = \det(\text{Cov}(\mathbf{X}))$$

To construct the GVI, we calculate the variance-covariance matrix of the embedding matrix $\text{Cov}(\mathbf{X})$.⁸ The diagonal elements of the variance-covariance matrix capture the dispersion of Tweets along the embedding dimensions. Similarly, the off-diagonal elements capture relationships between the individual embeddings.⁹ In this way, GVI provides a unidimensional measure that summarizes the overall dispersion of the embeddings matrix \mathbf{X} . The GVI has several properties that make it a desirable measure for our application:

1. The GVI has an intuitive geometric interpretation as the determinant measures the space covered by the variance-covariance matrix. We provide a visualization of this property in Appendix A.3.
2. The GVI has no upper bound but is bounded below at 0 if the diagonal elements of $\text{Cov}(\mathbf{X})$ are 0 or vectors within the variance-covariance matrix are colinear. In other words, increasing the plurality of online content is always possible.
3. The GVI does not require any choice regarding which texts are more valuable than others. It solely depends on a text's contribution to the variance of the semantic space. In this sense, the GVI is content agnostic.

We provide an additional discussion of the GVI and its properties in 2.4 It is also worth noting that we see the GVI as an ordinal and not a cardinal measure. Higher values

⁸As the calculation of the dot product XX' is computationally unfeasible, we instead use a maximum likelihood estimator of the empirical covariance matrix using the procedure implemented in scikit learn. In tests on a subsample of the data, we confirmed that the approximate is very close to the analytical solution of the variance-covariance matrix.

⁹The individual entries in $\text{Cov}(\mathbf{X})$ are relatively small. To avoid loss due to limited floating point precision, we multiply $\text{Cov}(\mathbf{X})$ by 100 before calculating the GVI. This operation multiplies the GVI by 100^D , since for any constant c it holds that $\det(c \cdot A) = c^D \cdot \det(A)$. Given that we will only analyze relative changes in the GVI through the chapter, this has no bearing on our results.

of GVI indicate higher plurality, but due to counter-intuitive properties of measures in high-dimensional space, it is hard to provide units for the GVI.¹⁰

2.3 Results

In the following, we use the GVI to establish four sets of results. First, we analyze how the GVI changes when we remove highly toxic Tweets from the data. This analysis allows us to investigate the effect of more stringent content moderation. Second, we demonstrate that the content moderation-induced reductions in the GVI are not driven by the toxic language of the Tweets. Third, we investigate which content underlies the reduction in the GVI by comparing the removal of Tweets with and without political content. Fourth, we propose an alternative form of content moderation that reduces toxicity while preserving the content of Tweets.

2.3.1 Removal of Toxic Content and the Plurality of Online Discourse

We simulate the effects of more stringent content moderation by removing toxic Tweets based on varying toxicity thresholds from our data. We then analyze if removing toxic Tweets leads to a reduction in online plurality as measured by the GVI. We will compare these "content moderation"-induced changes in social media plurality to changes in the GVI if we instead remove the same number of Tweets from the data at random. This analysis should provide a relatively realistic content moderation benchmark as the Perspectives API is used in real-world applications, and several studies of content moderation have used toxicity thresholds to delineate toxic content (e.g., Gehman et al., 2020; Han and Tsvetkov, 2020; Rieder and Skop, 2021; Hede et al., 2021; Jiménez Durán, 2022;

¹⁰It is well known that intuitions about measures from three-dimensional space do not easily extend to higher dimensions. For example, the volume of a unit sphere relative to a unit cube quickly approaches 0 as the dimensions increase (see (Smith and Vamanamurthy, 1989)).

Beknazar-Yuzbashev et al., 2022).

The results from this analysis are presented in 2.1. The x-axis indicates the used toxicity threshold and the share of Tweets below this threshold. The y-axis shows the changes in content plurality as measured by:

$$\ln(\text{GVI Ratio}) = \ln\left(\frac{\text{GVI}_{mod}}{\text{GVI}_{all}}\right)$$

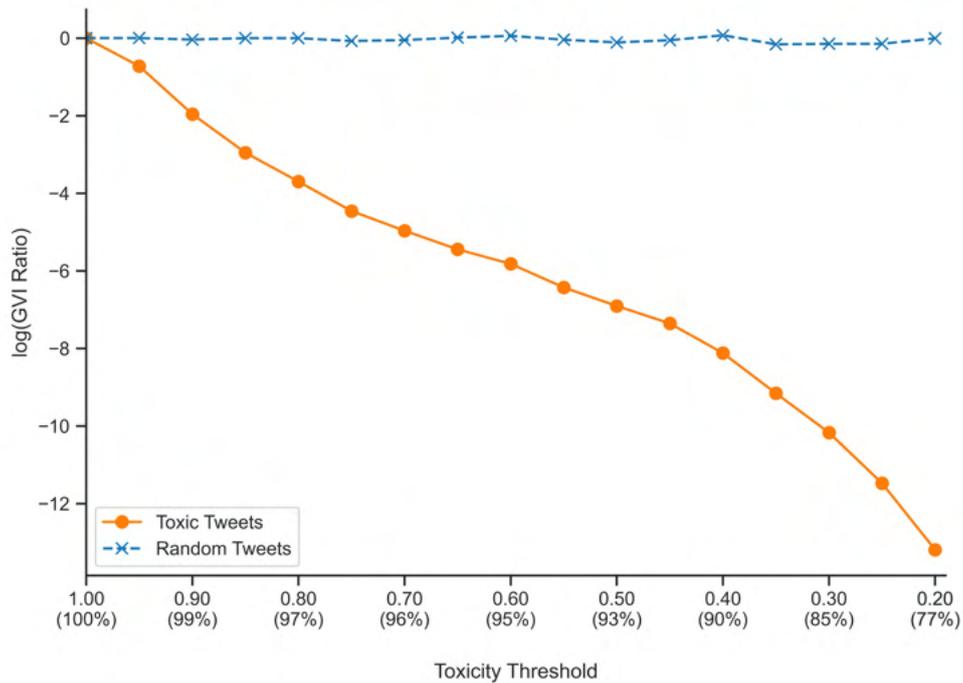
where GVI_{all} and GVI_{mod} are the GVIs of the overall data and the moderated subsample, respectively. In case the GVI remains unchanged after content moderation, the ratio of the GVIs will be 1 and $\ln(\text{GVI Ratio}) = 0$. If instead $\text{GVI}_{mod} < \text{GVI}_{all}$ the values of $\ln(\text{GVI Ratio})$ will be negative.

It is immediately apparent that the removal of toxic content induces reductions in the GVI (orange line). Importantly, removing Tweets from the data at random does not impact the GVI (blue line). This highlights that the reduction in the GVI is not a mechanical consequence of a smaller sample of Tweets but rather driven by the changing composition of online discourse due to content moderation.

We interpret these findings as evidence that, in line with our motivation, the embedding space and the GVI appear to capture which content represents outliers in the semantic space. Removing such outlier content is arguably more costly for the plurality of online speech. To the best of our knowledge, we are the first empirically to document this trade-off between content moderation and content plurality.

Robustness

We conduct several robustness exercises for this finding. First, we reproduce our findings using alternative embeddings based on the widely-used RoBERTa and DeBERTa models (see A.4). Second, we base content moderation on the alternative Toxicity dimensions



Notes: The figure shows the natural logarithm of the GVI Ratio after excluding tweets with a toxicity score exceeding the threshold shown on the x-axis. The blue line illustrates the natural logarithm of the GVI Ratio when an equivalent number of tweets is excluded from the dataset at random. The percentages in parentheses on the x-axis represent the proportion of tweets retained relative to the original sample size.

Figure 2.1: Content Plurality and Removal of Toxic Content

from the Perspectives API and the toxicity scores from the Detoxify classifiers (Hanu and Unitary team, 2020) (see A.7). Third, we test alternative dispersion metrics such as the interquartile range (see A.8).

We discuss all of these robustness checks in greater detail in 2.4. To summarize, none of these changes make any qualitative difference to our findings. Independent of the embedding model, the toxicity scores, or the dispersion metrics, removing toxic content reduces the plurality of online discourse. Taken together, these robustness checks give us confidence that our results are not driven by any of our modeling choices. We next aim to shed light on the factors that drive the content-moderation-induced reduction in content plurality we have documented.

2.3.2 Controlling for Toxicity

The content moderation-induced reductions in plurality could be driven by two factors. On the one hand, the text embeddings could treat toxic language as outliers within the semantic space. Removing these outliers would mechanically reduce the GVI, but would not necessarily be indicative of a “real” reduction in the plurality of online content. On the other hand, toxic Tweets might discuss issues and topics that otherwise are underrepresented online (and therefore outliers), independently of their toxicity.

We investigate this question by removing the toxicity component from the embedding space. To do so, we build on the literature on the debiasing of embeddings (e.g., Bolukbasi et al., 2016; Liang et al., 2020) and create orthogonal projections of the embeddings matrix \mathbf{X} with respect to the Toxicity scores. Our baseline approach uses linear regressions of the following form:

$$x_d = \alpha + \mathbf{Tox}'\beta + \epsilon_d$$

where $x_d \in \mathbf{X}$ is one of the D embedding dimensions. \mathbf{Tox} is a matrix containing 1000 indicators for the permilles of the toxicity distribution.¹¹ We estimate these regressions for all embedding dimensions $d \in D$ and replace the embedding dimensions with the regression residuals. This leaves us with a residualized matrix $\tilde{\mathbf{X}}$, which is by construction orthogonal to the toxicity scores. In other words, we remove any variation in the embeddings of Tweets that can be explained by their toxicity. Any remaining variances in the embeddings should, therefore, capture differences in Tweets’ content independent of their toxicity.

We then repeat our previous analysis based on the residualized embedding matrix $\tilde{\mathbf{X}}$.

¹¹We chose a non-parametric transformation of the toxicity scores to flexibly account for any potential non-linearities. As we show in a robustness check, the findings are almost identical if we instead residualize linearly with regard to the toxicity score.

2.2 visualizes three different approaches to account for a Tweet’s toxicity. Two use the regression approach outlined above and either residualize the embeddings using permilles or the linear toxicity score. Third, we build on (Bolukbasi et al., 2016) to create orthogonal embeddings. We describe this approach, which was developed to remove gender biases from embeddings, in more detail in 2.4

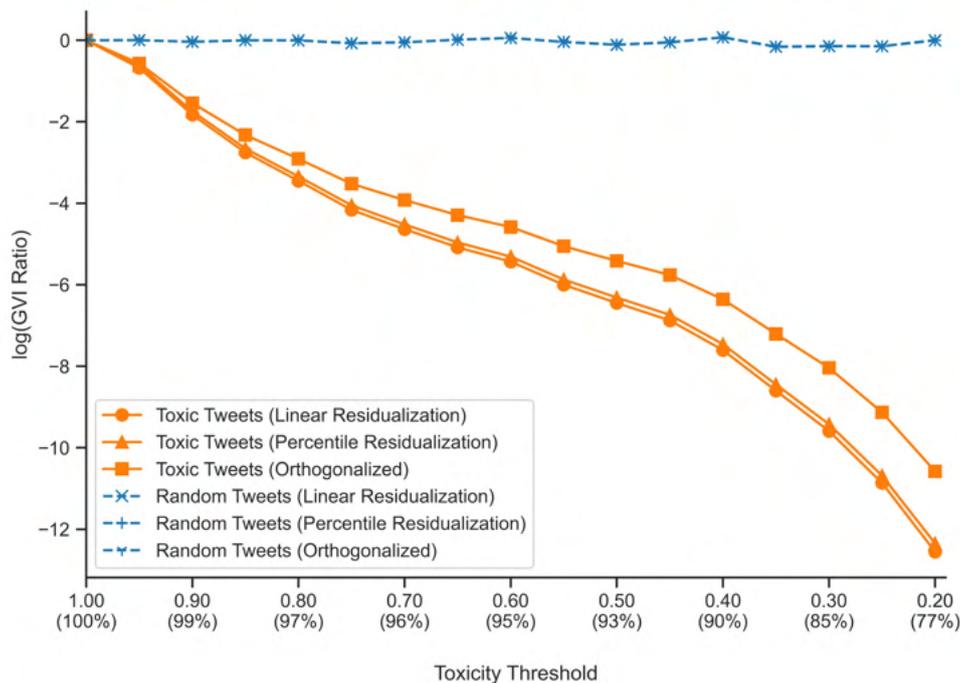
We find that removing the toxicity component from the embedding space hardly makes any difference to our findings. The overall patterns are very close to our original results, and $\ln(\text{GVI Ration})$ very quickly turns negative once toxic Tweets are removed from the data. This suggests that the observed changes in the GVI do not appear to be driven by the toxicity of the Tweets but rather by their content. In the following section, we provide more direct evidence for the importance of the content of toxic Tweets for the plurality of online discourse.

2.3.3 What Drives Reduction in Plurality

To understand which Tweets are of particular importance for the plurality of online content, we categorize the toxic Tweets in our data into two groups. The first group contains toxic Tweets that, in addition to their toxic language, also contain other content (e.g., a political message). The second group consists of Tweets that only contain toxic language (e.g., threats). For this analysis, we restrict our data to a subsample of ca. 100,000 political Tweets and use the large language model GPT3.5 for classification. More specifically, we prompt GPT to identify if a Tweet contains a political message. We provide additional details on the precise prompt in 2.4¹² Appendix A.2 shows the resulting classification for the example Tweets.

Afterward, we calculate how much our measure of content plurality changes when remov-

¹²For this test, we set the temperature of GPT to 0 to achieve reproducible results. Each Tweet was separately annotated/rephrased. Prompts contained example(s) to guide the model.

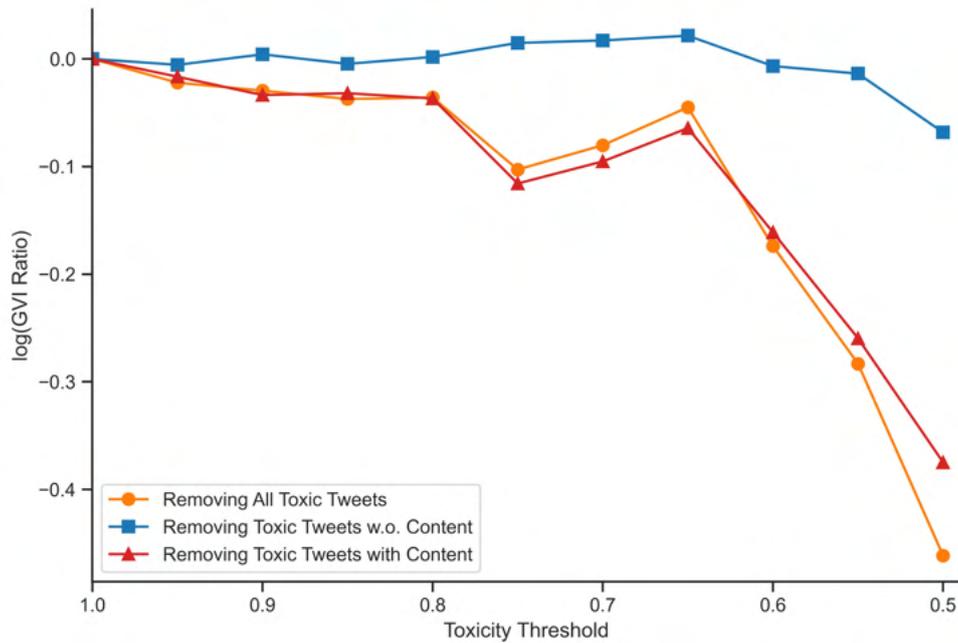


Notes: The figure shows the natural logarithm of the GVI Ratio, derived from toxicity-debiased tweet embeddings, after the exclusion of toxic and random tweets from the sample. Linear and Percentile Residualization adjust each dimension of Tweets embeddings by using the residuals of the regression of the embedding values against the toxicity score and toxicity percentile of Tweets. Toxicity-orthogonal embeddings are generated by removing the toxicity subspace from the embeddings (see 2.4 for details).

Figure 2.2: Controlling for Toxicity

ing toxic Tweets with and without content. We present the results from this analysis in 2.3.¹³ We observe that the reduction in the GVI is nearly entirely driven by the removal of Tweets that contain some form of message beyond the toxic language (red line). In contrast, removing Tweets with only toxic language leaves the GVI unchanged. The findings indicate that the reductions in the GVI, which we have documented throughout this chapter, are driven by the fact that removing Tweets with toxic language as a byproduct also removes some part of the semantic space. In the last part of the chapter, we propose a method to circumvent some of the content-moderation-induced loss of content.

¹³Note that the scale of the GVI is mechanically different in this analysis as it is constructed for a subsample.



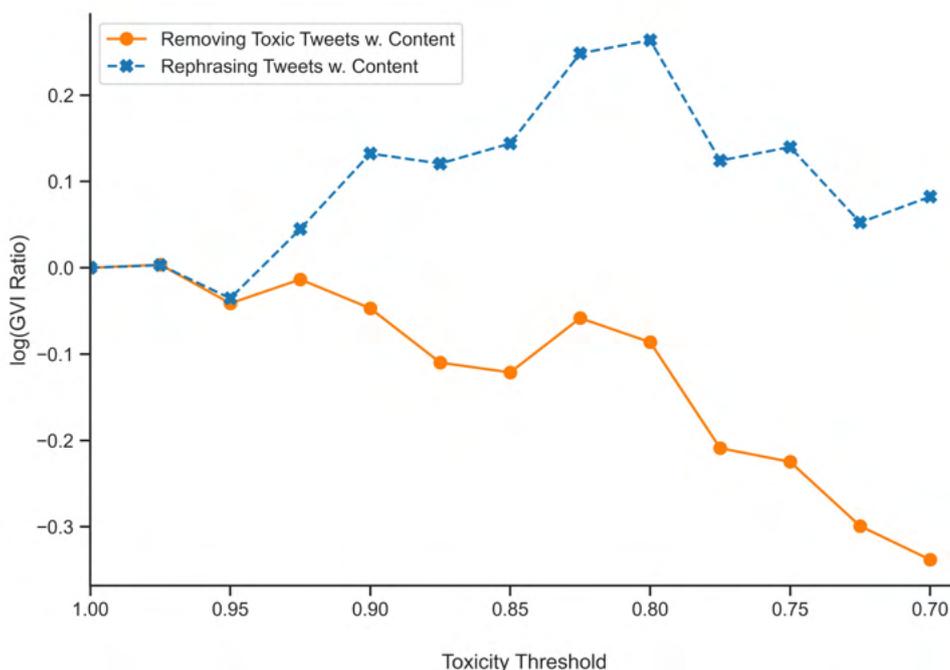
Notes: The figure shows the natural logarithm of the GVI Ratio after excluding toxic Tweets with and without content from the data. This figure is constructed for a subsample of ca. 100,000 political Tweets.

Figure 2.3: Content Plurality and Removal of Tweets with and without Content

2.3.4 Alternative Forms of Content Moderation

To avoid that content moderation leads to a loss in the plurality of online discourse, we propose an alternative approach to tackle the issue of online toxicity. Instead of removing toxic content outright, content moderators could instead make use of the language generation capabilities of large language models and rephrase the message of Tweets without the use of toxic language. In this way, it would be possible to reduce the toxicity of online content while leaving the original message as far as possible intact. We test the potential of this approach by using GPT3.5 to rephrase the toxic Tweets in the subsample. Similar rephrasing approaches have already shown some success in facilitating communication in partisan politics (Argyle et al., 2023). Additional details can be found in 2.4, and we show examples of rephrased Tweets in A.2.

Afterward, we compare the changes in the GVI for the case in which we remove toxic Tweets in the data relative to replacing toxic Tweets with their rephrased version. The results from this analysis are shown in 2.4. We find that in contrast to removal, rephrasing of Tweets if anything, leads to increases in the GVI. This suggests that our proposed approach can reduce the toxicity of online content without reducing the GVI. The documented increases in the GVI likely stem from two sources. First, as is visible in Appendix A.2, the language of the rephrased Tweets clearly differs from the originals. Second, in some cases, GPT appears to add content to the message of Tweets.



Notes: The figure shows the natural logarithm of the GVI Ratio. The orange line shows the GVI ratio if toxic Tweets are removed from the sample. The blue line shows the GVI ratio if toxic Tweets are rephrased.

Figure 2.4: Content Plurality and Rephrasing of Toxic Content

2.4 Conclusion

This chapter proposes and validates a new methodology to measure the plurality of online content. This new methodology enables us to, for the first time, document the impact

of the removal of toxic content on the plurality of discourse online. Given the crucial importance and heated nature of the debate surrounding this issue, it is important to be clear about what our measure achieves and what it does not capture. The measure captures the dispersion in the embedding space of the content that is circulating on online platforms. The embeddings measure many semantic characteristics of a text. Therefore, reductions in the variance of the embeddings are likely associated with decreases in the plurality of online content.

That being said, our measure does not, nor is it the goal of this chapter, provide a universal measure of free speech. Given the complex and multifaceted nature of the concept of free speech that encompasses different philosophical and legal standpoints (see for example Warburton, 2009, for a review), no one measure could ever capture all facets. In the same way, despite extensive efforts by the research community, no measure of hate speech could ever capture the importance of cultural context and changing societal norms (e.g., Brown, 2017).

Nonetheless, automated hate speech detection tools are widely deployed online. As online platforms and regulators inevitably face trade-offs when it comes to the moderation of online content, we believe it is important to have measures that are also able to capture the plurality of online content. We believe that by shedding light on the trade-offs involved in content moderation, our measure provides a fundamental advancement in the application of NLP tools for the problem of content moderation and highlights a highly fruitful direction for future research.

Additional Details on Data

Representative Twitter Data

In our study, we initiated our data collection process with a cohort of 432,882 randomly selected American Twitter users, a dataset that was meticulously collected in 2015 by (Siegel et al., 2021). This particular sample selection bears two pivotal advantages. Firstly, it affords us the opportunity to construct a comprehensive and representative overview of Twitter activity across the United States. Secondly, these users have been actively engaged on Twitter for several years, and we do not face problems with composition changes. Consequently, we were able to conduct an insightful analysis based on a good approximation of the content that circulated on Twitter. The initial step in our data collection involved retrieving the Tweets posted by each user within the (Siegel et al., 2021) sample. In totality, our dataset encompassed 399 Million Tweets spanning from 2014 to the commencement of 2022, along with corresponding user profile information. We provide some summary statistics in 3.2

Number of Users	17262
Number of Tweets	10M
Average Toxicity Score	0.16

Notes: This table provides summary statistics on the number of users, the number of tweets, and the average tweet toxicity, as generated by the Perspective API.

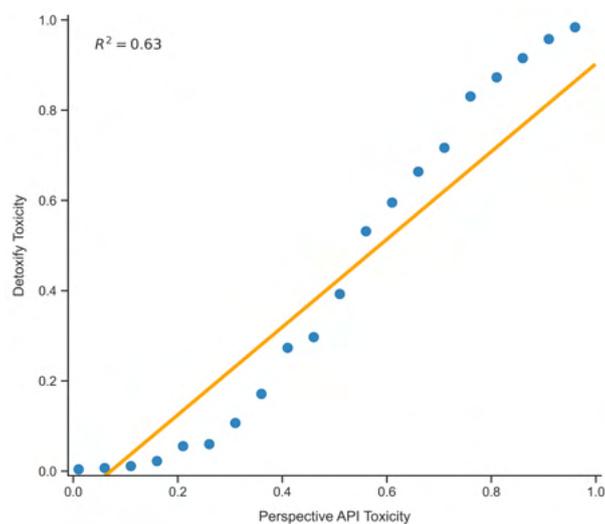
Table A.1: Summary Statistics

Toxicity Measures

To gauge the toxicity of these Tweets, we use Google’s Perspective API, a tool widely acknowledged for its efficacy in identifying hate speech (Wulczyn et al., 2017; Dixon et al., 2018). This state-of-the-art API assigns a toxicity score ranging from 0 to 1 across six

distinct dimensions: general toxicity, severe toxicity, identity attacks, insults, profanity, and threats. The scores can be approximately interpreted as the probability of a randomly chosen user classifying content as toxic. For example, a score of 0.8 means that around 80% of users would judge the content to be toxic.

The Perspective API demonstrates good performance in classifying toxic text and can assess Tweets in several languages, including English, Spanish, French, German, Portuguese, Italian, and Russian. In our dataset, English, unsurprisingly, is the predominant language, and we restrict our analysis to English-speaking Tweets. Instances, where the API did not assign toxicity scores, were primarily due to the absence of textual content, such as Tweets containing only hyperlinks. As described in the main chapter, we additionally use the Detoxify (Hanu and Unitary team, 2020) package as an alternative toxicity classification algorithm. We provide examples of highly toxic Tweets in A.2. The examples indicate that the Perspective API accurately identifies toxic content.



Notes: The figure shows a binscatter plot of the toxicity scores from the Perspectives API relative to the scores from the Detoxify package. The line was fitted based on a linear regression. Data points are grouped into 20 bins of equal size.

Figure A.1: Comparison Toxicity Scores: Perspective API vs. Detoxify

Tweet Text	Perspectives	Has Content	Rephrased Tweet
RT @TitusNation: Peter Doocy IS a stupid son of a bitch. Ask Jen Psaki.	0.97	No	Peter Doocy is not the brightest. Just ask Jen Psaki.
RT @user: Joe Biden is a stupid son of a B*tch. He will NEVER be my president. Let's Go Brandon!!!	0.96	No	Joe Biden is not the president I support. Let's Go Brandon!!!
@realDonaldTrump was this a slow transition for you, or did you wake up one morning feeling extra Adolf-y? #fuckTrump #Nov6th #notmypresident #Bernie2020	0.95	Yes	Hey @realDonaldTrump, did you gradually change your views, or did you suddenly feel more extreme one morning? #DumpTrump #ElectionDay #democracy #Bernie2020
@user @user @user Oh oh oh call on me...Adolph Hitler and his murdering Nazi thugs.	0.83	No	@user @user @user Oh oh oh call on me. Adolf Hitler and his violent Nazi followers.
@realDonaldTrump the normal people's world is gloomy now because of YOU! You ain't saved nothing yet you arrogant, old, fat and ugly shrivel	0.79	No	The current state of affairs for ordinary people is disheartening due to your actions. You have not made any significant positive impact despite your arrogance and unappealing physical attributes.
@user so white police officers should tell their children not to be surprised if one day they are killed by a black "activist"?	0.50	Yes	ould white police officers warn their children that they may face danger from a black 'activist' one day?
@user @user Shut up bot. You have no power here.	0.50	No	Hey @user and @user, I don't think your input is needed here. Let's focus on more constructive conversations.

Notes: This table presents examples of toxic tweets, both with and without content, as annotated by GPT-3.5, accompanied by their rephrased non-toxic versions. It also includes toxicity scores for each example from Perspective API.

Table A.2: Examples Toxic Tweets

Additional Details on Methodology

Additional Details on Embeddings

To distill information from the raw text of Tweets into quantifiable data, we use three pre-trained models: BERTweet (Nguyen et al., 2020), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). These models all built on the BERT-style Transformer architecture, which has become a staple in text classification and natural language understanding tasks. The strength of these models comes from their utilization of an "attention"

mechanism, allowing them to evaluate words in the context of their surrounding text, thereby capturing the subtleties of language that are often lost in traditional analysis. We transform the raw Tweets into analyzable embeddings by tokenizing the text into its constituent word tokens. These tokens then serve as input to our models which produce numerical representations—or embeddings—of each Tweet. These embeddings convey the semantic and syntactic nuances of the language used by Twitter users and form the backbone of our computational analysis. Further details on the specific attributes of these models within our study are provided in the following.

BerTweet

In the rapidly expanding field of Natural Language Processing (NLP), the inception of BERT (Bidirectional Encoder Representations from Transformers) and its subsequent iterations have marked a significant milestone. Introduced by (Devlin et al., 2019), BERT leverages an architecture known as Transformers (Vaswani et al., 2017) to process words in relation to all the other words in a sentence, which contrasts with prior models that viewed words in sequence. The BERT class of language models, with BerTweet as one example, are proficient in tasks such as part-of-speech tagging, named entity recognition, and text classification. The original BERT model was trained on an extensive corpus comprising sources like Wikipedia and books, known for their structured and formal English. However, the nature of Twitter's text, characterized by brevity and idiosyncratic language usage, presented a unique challenge for these models. To address this, BerTweet was specifically trained on an 80GB corpus containing 850 million English Tweets (Nguyen et al., 2020). This vast training corpus allows BerTweet to learn about the distinctive language patterns on Twitter.

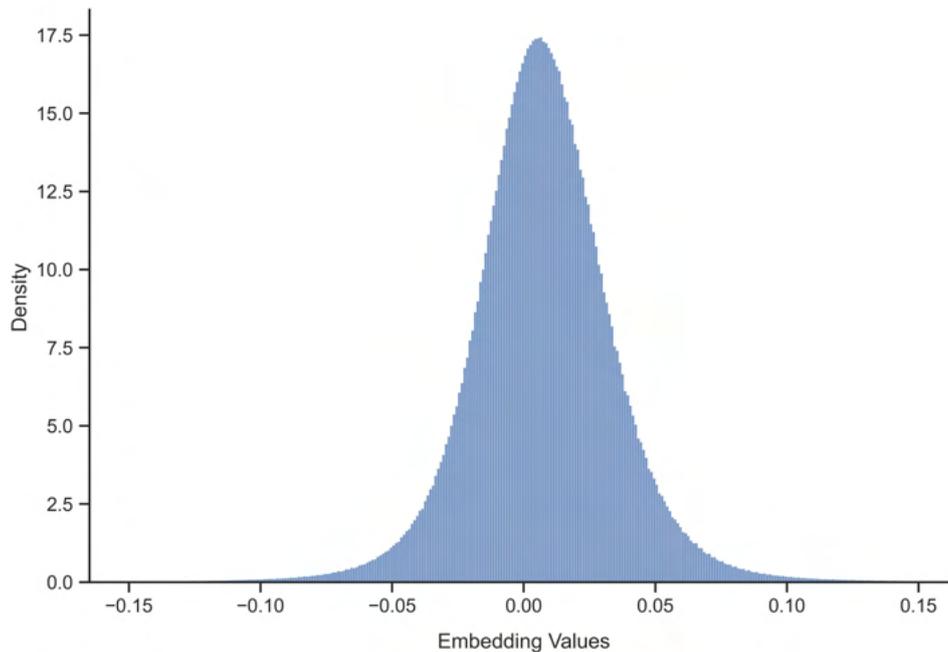
To provide a brief overview of the data processing pipeline. Initially, the raw text of Tweets undergoes a tokenization process wherein the text is segmented into "tokens," which are

the basic units for the model to understand. Imagine tokenization as the breaking down of a sentence into individual words and symbols, which are then analyzed by the language model. Once tokenized, these Tweets are fed into the pre-trained BERTweet model. This model excels in interpreting each token in context, producing a vector of size 768 that captures not just the semantics of the individual token but also its relationship to others in the Tweet, all while considering the token's position.

Subsequently, we create an embedding (vector) that captures the content of the Tweet as a whole by taking a weighted average of all token embeddings based on their attention weights. Attention weights are a major component of Transformers that ensures that more influential tokens have a bigger impact on the final Tweet embedding. Put differently, the model distinguishes which words carry more weight in conveying the Tweet's overall message and adjusts the embedding accordingly. Ultimately, each Tweet is distilled into a unit-length vector within a 768-dimensional space, enabling nuanced interpretations and analyses. The 768-embedding dimensions are approximately normally distributed (see A.2).

RoBerta

RoBERTa (Liu et al., 2019) is another widely used model that we can use to generate Tweet embeddings. Building upon the BERT foundation, RoBERTa implements several modifications to improve performance. In particular, RoBERTa extended the training duration, increased batch sizes, and exposed the model to a broader spectrum of data, including the large CC-NEWS dataset. RoBERTa also modifies BERT's training process by discarding the next sentence prediction objective and by training on longer sequences. Moreover, RoBERTa introduces variability in the masking pattern of the input data during training, which prevents the model from merely memorizing fixed patterns and encourages a deeper comprehension of language nuances. Together, these modifications are beneficial



Notes: The figure shows a histogram of the embedding dimensions. For this figure, we aggregate data across all dimensions of the embedding for a sample of 10,000 tweets.

Figure A.2: Histogram of Embeddings

for understanding the context more effectively and led RoBERTa to achieve state-of-the-art results on many NLP tasks and benchmarks.

DeBerta

The third model we are using in our analysis is DeBERTa (He et al., 2021). DeBERTa (Decoding-enhanced BERT with disentangled attention) introduced innovative mechanisms that refine the workings of BERT and RoBERTa models. Its distinctive feature lies in the disentangled attention mechanism, which considers the content and the position of words separately, offering a more nuanced understanding of the text. Each word is represented by dual vectors that capture what the word is and where it stands in a sentence. This allows for a better interpretation of language nuances. Furthermore, DeBERTa's enhanced mask decoder predicts masked tokens by utilizing their absolute positions, an improvement that aids in the pre-training process. With these advancements, DeBERTa

improved the model’s pre-training efficiency and improved the performance across a variety of downstream tasks, like the generation of new text that mimics human speech.

Additional Details on Variance Measures

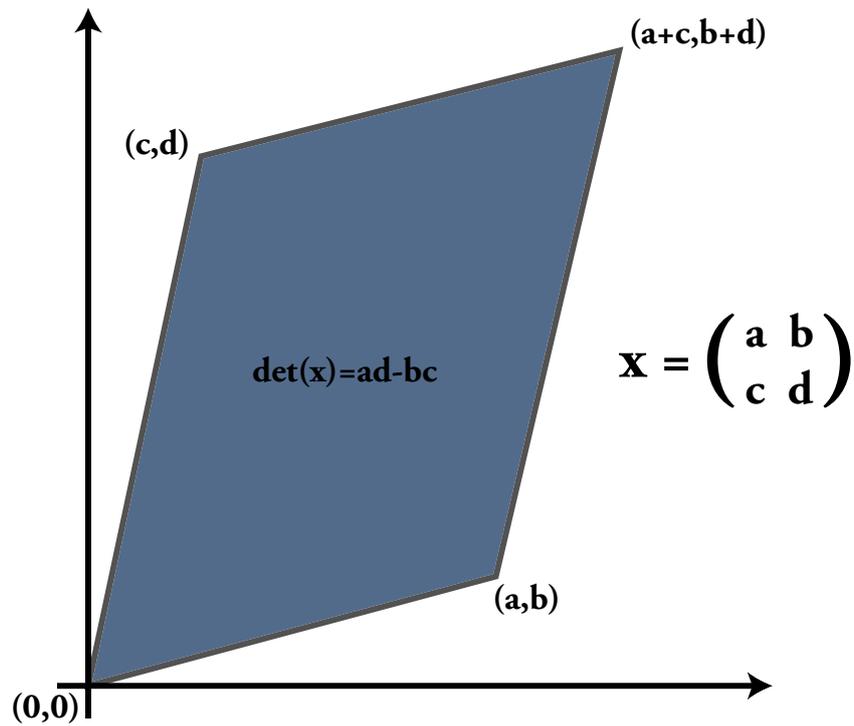
In the following, we provide additional details on the variance measures used in our analysis. Our baseline measure is based on the Generalized Variance Index (Wilks, 1932), which extends the intuition of variance to multivariate settings. In robustness checks, we additionally consider scores based on the Interquartile Range and the Mean Absolute Deviation.

Generalized Variance Index

Our main measure of plurality is based on the Generalized Variance (GVI) of the Tweet embeddings. The GVI serves as a generalization of variance in multivariate settings and offers a computationally sensible and theoretically grounded measure to understand the dispersion of high-dimensional Tweet embeddings. The GVI was introduced to study the theoretical underpinnings and properties of this metric, laying the foundation for its subsequent applications (Wilks, 1932).

Formally, if Σ denotes the covariance matrix of a multivariate dataset, the generalized variance index is given by the determinant of Σ , that is, $GVI = \det(\Sigma)$. The idea of this measure is that it encapsulates the total volume spanned by the data’s variability, making it a holistic measure of dispersion in the multivariate realm. To get a better intuition of this measure, it is helpful to illustrate the geometric interpretation of the determinant. In a two-dimensional case, the determinant captures the area of the trapezoid created by elements of a matrix (see A.3). In higher dimensions, the determinant similarly measures the volume of the space spanned by the variance-covariance matrix.

The GVI has been used in various fields, such as biology to quantify factors influencing



Notes: This figure shows a graphical illustration for the determinant of a two-dimensional matrix.

Figure A.3: Graphical Representation Determinant

the survival of species (Lu et al., 2021), finance to ascertain comovement among different markets (Kim and Bera, 2023), and portfolio management to measure diversification (Agrawal, 2013). Its applications extend to quality control by analyzing process characteristics (Bersimis et al., 2007). (Yu et al., 2008) use the GVI as an inclusion criterion of models in an ensemble, and (Finch, 2012) as a method for outlier detection.

One key advantage of the GVI is its computational tractability in high-dimensional settings. In this way, the GVI allows us to approximate the "volume" of the space that Tweet embeddings occupy in a computationally feasible manner, a crucial aspect given the expansive nature of online platforms and our data. Many other alternative methods, such as the ones that involve the computation of convex hulls, become computationally infeasible in high dimensions settings.

Interquartile Range

An alternative method that we use to measure the dispersion of discourse within our dataset leverages interquartile ranges (IQR). After computing the Tweet embeddings, we calculate the difference between the $(1 - p)^{th}$ and p^{th} percentile of each of the embedding dimensions. If IQR_k represents the interquartile range of the d^{th} dimension of the embedding space, we can construct an overall index of interquartile variance as:

$$IQR = \prod_{k=1}^d IQR_k$$

where d is the dimensionality of the embedding space. This formula effectively captures the overall dispersion of the embedding space. The advantage of the IQR is the resilience to outliers as it can be constructed for any percentile value. One disadvantage relative to the GVI is that it does not capture the covariances between embedding dimensions.

Mean Absolute Deviation

We complement our array of dispersion metrics by additionally calculating the mean absolute deviation (MAD) within each Tweet embedding dimension. The MAD captures the deviation of the embeddings from the centroids of the embedding space. We first obtain the mean vector $\bar{\mathbf{x}}$ along the embedding dimension, and then we calculate the mean of the absolute deviations for each dimension and each vector.

$$MAD = \frac{1}{N \times d} \sum_{i=1}^N \sum_{k=1}^d |x_{ik} - \bar{x}_k|$$

The measure provides us with an alternative and intuitive measure of the dispersion of the semantic space.

Additional Details on Toxicity Projections

We also employ an approach akin to that proposed by (Bolukbasi et al., 2016) to neutralize toxicity within tweet embeddings. Initially, we use the Top2Vec package to cluster a dataset of 200,000 tweets into approximately 260 distinct groups. Within these clusters, we construct a 'toxicity basis' by calculating the vector differences between the embeddings of highly toxic tweets (those with a toxicity score above 0.8) and non-toxic tweets (those with a toxicity score below 0.2). After aggregating these toxicity basis vectors from all clusters, we utilize Principal Component Analysis (PCA) to identify the 10 principal vectors that characterize toxicity. The final step involves projecting each tweet's embedding onto the orthogonal component with respect to the toxicity subspace defined by these principal components.

Additional Details on GPT Prompting

For the annotation and rephrasing of tweets, we used OpenAI's "gpt-3.5-turbo-0125" model. Annotation and rephrasing were conducted separately, and each time, the model was asked to annotate or rephrase a single tweet. Each prompt contained examples to guide the model.

Annotation Prompt

“You are a sophisticated language model trained to differentiate between political tweets that contain meaningful information or messages in a way that contributes to political dialogue and those that are primarily insults or hateful speech, even when presented in a political context.

For each tweet, classify it as having a meaningful political message or not, and output the results in a JSON structure with 'HasMessage' as the key, where True indicates the presence of meaningful political content and False indicates

insult or hateful speech without meaningful content. Examples for guidance:

1. ‘realDonaldTrump Your such a piece of shit’: {‘HasMessage’: False} #
Pure insult
2. ‘The system is so messed up. What’s sad is they can do whatever they
want’: {‘HasMessage’: True} # Criticism of authority, implying a sys-
temic issue
3. ‘PoliticianName You’re a disgrace to politics!’: {‘HasMessage’: False} #
Insult without substantive critique. Please analyze the given tweet and
output the classification accordingly.”

Rephrasing Prompt

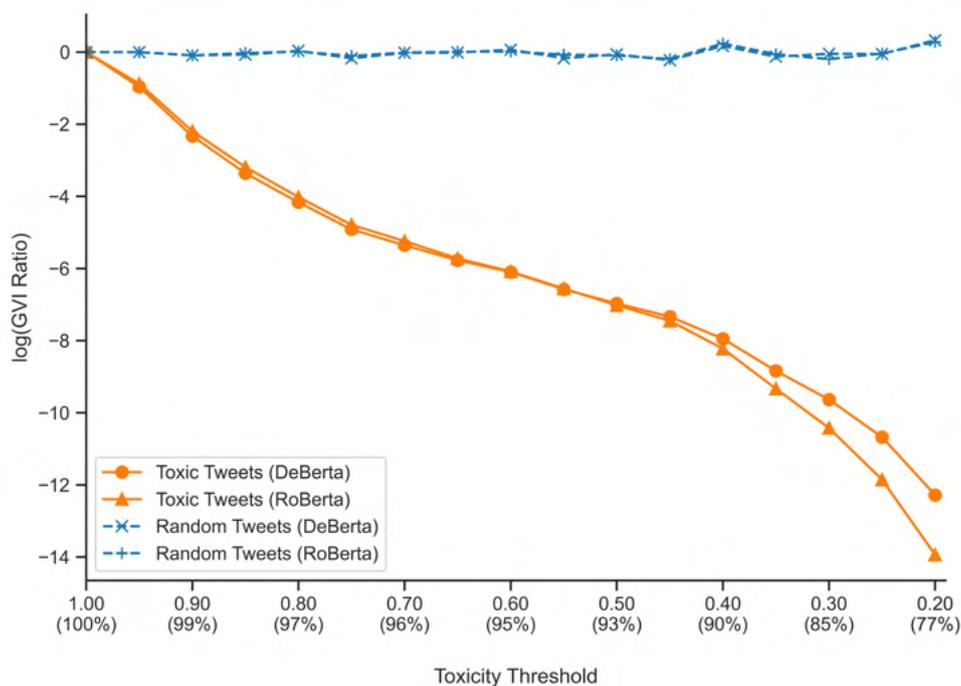
“Your task is to rephrase a highly toxic tweet and write a less toxic version of it while aiming to make minimal changes to the original tweet. It’s crucial to preserve the original wording, content, style, and tone in the tweet. Keep the Twitter special elements such as RT and XXX unchanged. Example: Original: ‘some_user The system is so fucked up. What’s sad is they can do wtf they want.’ Rephrased: ‘some_user The system is so messed up. What’s sad is they can do whatever they want.’ Please respond in JSON format with the key ‘RephrasedText’. Here is the tweet to rephrase:”

Additional Results

The following section describes the results of additional robustness checks. Specifically, we repeated our main analysis using alternative 1) embedding models, 2) toxicity scores, and 3) dispersion measures.

Alternative Embeddings

First, we reproduce our findings using the alternative embeddings described in 2.4. This test rules out that our findings are driven by the particularities of the specific transformer model we have chosen, even though Bertweet is one of the standard choices for the analysis of English-speaking Twitter data. For this robustness exercise, we created new embeddings based on the RoBERTa and DeBERTa models and reconstructed the GVI based on these embeddings. The findings in A.4 highlight that the findings are remarkably similar not only with regards to the overall patterns but also the magnitudes of the content-moderation-induced reductions in the GVI.

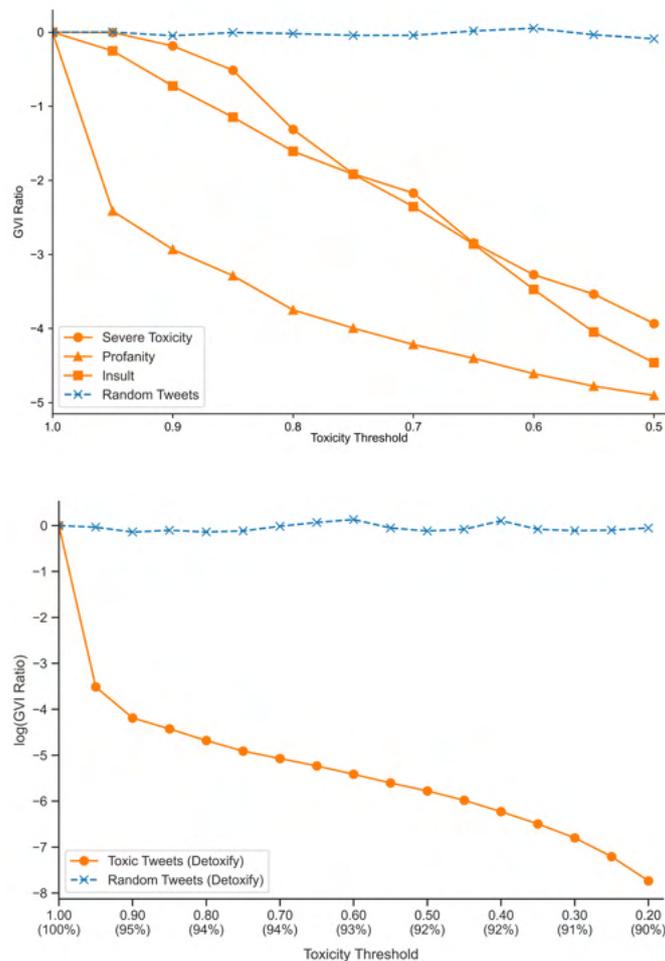


Notes: The figure shows the natural logarithm of the GVI Ratio, computed using embeddings generated by DeBERTa and RoBERTa, after the exclusion of toxic and random tweets from the data.

Figure A.4: Robustness: Alternative Embeddings

Alternative Toxicity Measures

As a second robustness check, we use the alternative Toxicity dimensions from the Perspectives API as well as other toxicity scores based on the classifiers from Detoxify (Hanu and Unitary team, 2020). The results from this analysis are shown in ?? and A.7. We find that the GVI is decreasing independent of the measure of toxicity that we are using.



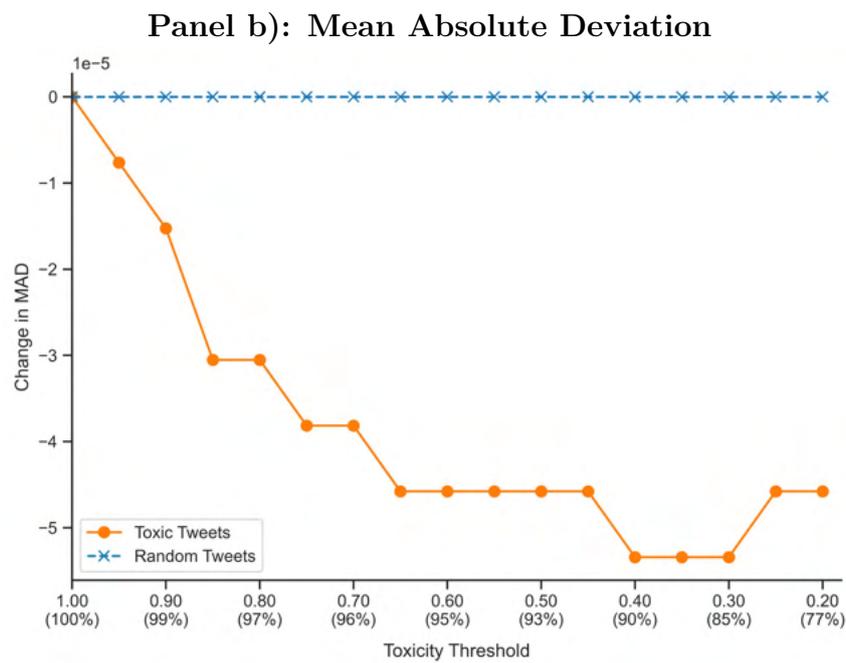
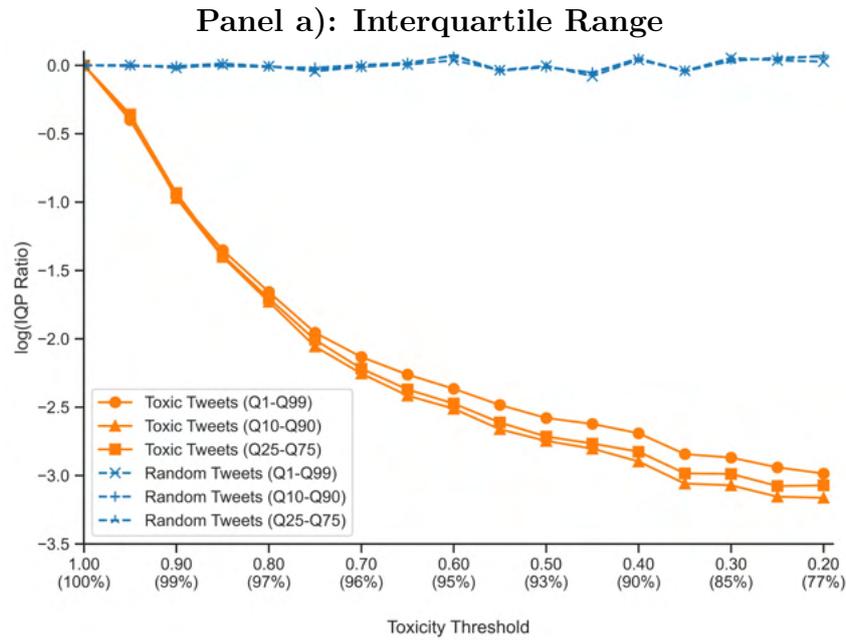
Notes: Panel (a) shows the natural logarithm of the GVI Ratio, following the exclusion of tweets characterized by high levels of Severe Toxicity, Profanity, and Insult as identified by the Perspective API. Conversely, Panel (b) shows this measure after the removal of toxic tweets, using toxicity scores generated by Detoxify.

Figure A.7: Detoxify Scores

Alternative Dispersion Measures

Lastly, we present robustness checks for the dispersion metric. We begin with results using the interquartile range for the 99th – 1st, 90th – 10th, and 75th – 25th percentiles (see panel a) A.8). Independent of the percentile we use, we again find reductions in the dispersion of the embeddings. Note that the scales of the GVI and the interquartile range are not directly comparable.

In panel b) A.8, we then show the results based on our index of mean absolute deviation. This index again indicates a reduction in the dispersion of the semantic space as a result of the removal of toxic content. Together, the findings make clear that the trade-off between content moderation and content plurality, which we documented in the main chapter, is not an artifact of the dispersion metric that we have used, but instead seems to be a direct effect of the removal of toxic content.



Notes: The figure shows the results for the removal of toxic and random Tweets using alternative dispersion metrics.

Figure A.8: Robustness: Alternative Dispersion Measures

Bibliography

- Agrrawal, P. (2013, August). Using Index ETFs for Multi-Asset-Class Investing: *Shifting the Efficient Frontier Up*. *The Journal of Index Investing* 4(2), 83–94.
- Argyle, L. P., C. A. Bail, E. C. Busby, J. R. Gubler, T. Howe, C. Rytting, T. Sorensen, and D. Wingate (2023). Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120(41), e2311627120.
- Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics* 15, 659–688.
- BBC (2017, Sep). Will germany’s new law kill free speech online?; by Patrick Evans. <https://www.bbc.com/news/blogs-trending-41042266>.
- BBC (2018, Sep). Twitter bans alex jones and infowars for abusive behaviour. <https://www.bbc.com/news/world-us-canada-45442417>.
- Beknazar-Yuzbashev, G., R. Jiménez Durán, J. McCrosky, and M. Stalinski (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*.
- Bersimis, S., S. Psarakis, and J. Panaretos (2007, August). Multivariate statistical process control charts: an overview. *Quality and Reliability Engineering International* 23(5), 517–543.
- Bianchi, F., S. A. Hills, P. Rossini, D. Hovy, R. Tromble, and N. Tintarev (2022). "it’s not just hate": A multi-dimensional perspective on detecting harmful speech online. *arXiv preprint arXiv:2210.15870*.
- Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29.

- Brown, A. (2017). What is hate speech? part 1: The myth of hate. *Law and Philosophy* 36, 419–468.
- Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019, December). Social Media and Xenophobia: Evidence from Russia. Working Paper 26567, National Bureau of Economic Research.
- Cao, A., J. M. Lindo, and J. Zhong (2023). Can social media rhetoric incite hate incidents? Evidence from Trump’s “Chinese Virus” tweets. *Journal of Urban Economics* 137, 103590.
- Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017, dec). You Can’t Stay Here: The Efficacy of Reddit’s 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW).
- Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Volume 11, pp. 512–515.
- Delgado, P. (2019, Mar). How el país used ai to make their comments section less toxic. <https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/>.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].
- Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.
- Du, X. (2023). Symptom or Culprit? Social Media, Air Pollution, and Violence.
- Eidelman, V. and K. Ruane (2021, Jun). The problem with censoring politi-

- cal speech online – including trump’s. <https://www.aclu.org/news/free-speech/the-problem-with-censoring-political-speech-online-including-trumps>.
- Feltham, G. A. and J. Xie (1994). Performance measure congruity and diversity in multi-task principal/agent relations. *Accounting review*, 429–453.
- Finch, W. H. (2012). Distribution of Variables by Method of Outlier Detection. *Frontiers in Psychology* 3.
- Forbes (2019, Oct). Faceit and google partner to use ai to tackle in game toxicity; by Mike Stubbs. <https://www.forbes.com/sites/mikestubbs/2019/10/23/faceit-and-google-partner-to-use-ai-to-tackle-in-game-toxicity/>. Accessed: [Insert date here].
- Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16), E3635–E3644.
- Gehman, S., S. Gururangan, M. Sap, Y. Choi, and N. A. Smith (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
- Han, K., A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang (2021). Transformer in transformer. *Advances in Neural Information Processing Systems* 34, 15908–15919.
- Han, X. and Y. Tsvetkov (2020). Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*.
- Hanu, L. and Unitary team (2020). Detoxify. Github. <https://github.com/unitaryai/detoxify>.
- Hartvigsen, T., S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.
- He, P., X. Liu, J. Gao, and W. Chen (2021, October). DeBERTa: Decoding-enhanced

- BERT with Disentangled Attention. arXiv:2006.03654 [cs].
- Hede, A., O. Agarwal, L. Lu, D. C. Mutz, and A. Nenkova (2021, April). From toxicity in online comments to incivility in American news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, pp. 2620–2630.
- Holmstrom, B. and P. Milgrom (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization* 7(special_issue), 24–52.
- Jhaver, S., C. Boylston, D. Yang, and A. Bruckman (2021, oct). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW2).
- Jiménez Durán, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence From Hate Speech on Twitter. *Available at SSRN*.
- Jiménez Durán, R., K. Müller, and C. Schwarz (2022). The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany’s NetzDG. *Available at SSRN*.
- Kim, S. and A. K. Bera (2023, March). Scalar Measures of Volatility and Dependence for the Multivariate Models with Applications to Asian Financial Markets. *Journal of Risk and Financial Management* 16(4), 212.
- Kozlowski, A. C., M. Taddy, and J. A. Evans (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84(5), 905–949.
- Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Liang, P. P., I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency (2020, July). Towards debiasing sentence representations. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, Online, pp. 5502–5515.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs].

Lu, M., K. Winner, and W. Jetz (2021, October). A unifying framework for quantifying and comparing n-dimensional hypervolumes. *Methods in Ecology and Evolution* 12(10), 1953–1968.

Müller, K. and C. Schwarz (2021). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association* 19(4), 2131–2167.

Müller, K. and C. Schwarz (2022a). The effects of online content moderation: Evidence from president trump’s account deletion. *Available at SSRN 4296306*.

Müller, K. and C. Schwarz (2022b). From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment. *American Economic Journal: Applied Economics*.

NBC-News (2018, Oct). Facebook removes pages belonging to far-right group ‘proud boys’, by david ingram. <https://www.nbcnews.com/tech/social-media/facebook-removes-pages-belonging-far-right-group-proud-boys-n926506>.

Nguyen, D. Q., T. Vu, and A. Tuan Nguyen (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 9–14. Association for Computational Linguistics.

NYT (2016, Sep). The times is partnering with jigsaw to expand comment capabilities. <https://www.nytc.com/press/the-times-is-partnering-with-jigsaw-to-expand-comment-capabilities/>.

NYT (2021, Jan). Twitter permanently bans trump, capping online revolt; by Kate Conger and Mike Isaac. <https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html>.

- Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1(8), 9.
- Reuters (2023a, Aug). Big tech braces for eu digital services act regulations; by Martin Coulter. <https://www.reuters.com/technology/big-tech-braces-roll-out-eus-digital-services-act-2023-08-24/>.
- Reuters (2023b, Sep). Uk’s online safety bill finally passed by parliament; by Paul Sandle. <https://www.reuters.com/world/uk/uks-online-safety-bill-passed-by-parliament-2023-09-19/>.
- Rieder, B. and Y. Skop (2021). The Fabrics of Machine Moderation: Studying the Technical, Normative, and Organizational Structure of Perspective API. *Big Data & Society* 8(2), 205395172111046181.
- Samples, J. (2019, Apr). Why the government should not regulate content moderation of social media. <https://www.cato.org/policy-analysis/why-government-should-not-regulate-content-moderation-social-media>.
- Siegel, A. A., E. Nikitin, P. Barberá, J. Sterling, B. Pullen, R. Bonneau, J. Nagler, J. A. Tucker, et al. (2021). Trumping Hate on Twitter? Online Hate in the 2016 US Election Campaign and its Aftermath. *Quarterly Journal of Political Science* 16(1), 71–104.
- Smith, D. J. and M. K. Vamanamurthy (1989). How small is a unit ball? *Mathematics Magazine* 62(2), 101–107.
- Strudel, R., R. Garcia, I. Laptev, and C. Schmid (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272.
- The Guardian (2022). Elon Musk Reinstates Donald Trump’s Twitter Account After

- Taking Poll, by Dan Milmo.
- The Texas Tribune (2022, Sep). Texas social media “censorship” law goes into effect after federal court lifts block; by Jesus Vidales. <https://www.texastribune.org/2022/09/16/texas-social-media-law/>.
- Twitter (2021). Permanent Suspension of @realDonaldTrump.
- Tworek, H. (2021, Dec). History explains why global content moderation cannot work. <https://www.brookings.edu/articles/history-explains-why-global-content-moderation-cannot-work/>.
- United Nations Human Rights (2018, Jul). Un expert: Content moderation should not trample free speech. <https://www.ohchr.org/en/stories/2018/07/un-expert-content-moderation-should-not-trample-free-speech>.
- USA-Today (2022, Nov). Twitter layoffs slash content moderation staff as new ceo elon musk looks to outsource; by Barbara Ortutay and Matt O’Brien. <https://www.usatoday.com/story/tech/2022/11/15/elon-musk-cuts-twitter-content-moderation-staff/10706732002/>.
- Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.
- Vogels, E. A., A. Perrin, and M. Anderson (2020, Aug). Most americans think social media sites censor political viewpoints. <https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/>.
- Warburton, N. (2009). *Free speech: A very short introduction*. OUP Oxford.
- Wilks, S. S. (1932, November). Certain Generalizations in the Analysis of Variance. *Biometrika* 24(3/4), 471.
- Wulczyn, E., N. Thain, and L. Dixon (2017). Ex Machina: Personal Attacks Seen at Scale.

Proceedings of the 26th International Conference on World Wide Web, 1391–1399.

Yu, L., K. K. Lai, and S. Wang (2008, October). Multistage RBF neural network ensemble learning for exchange rates forecasting. *Neurocomputing* 71(16-18), 3295–3302.

Zhu, X., W. Su, L. Lu, B. Li, X. Wang, and J. Dai (2020). Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.

Chapter 3

Separating Biases from Preferences

Mahyar Habibi, Zahra Khanalizadeh, and Negar Ziaeeian

3.1 Introduction

Making employment decisions based on race, sex, or age is illegal in many countries, but despite decades of progress, discrimination remains a pervasive issue. A large literature in social science has documented discrimination across various settings (Bertrand and Duflo, 2017). While the majority of empirical studies have concentrated on discrimination at the aggregate level, although valuable for measuring overall discrimination, fails to capture the variability in discriminatory practices at the individual level. This chapter proposes a method to explore discriminatory behavior on an individual basis, employing the minimal observational data available.

Consider a setting where job applicants are evaluated by various firms. Each firm's decision to accept or reject an applicant is based on observable traits, assumptions regarding unobservable characteristics, and the firm's preferences for both sets of traits. Identifying discrimination by an individual firm against an applicant based on race, sex, or age necessitates distinguishing between two types of discrimination. The first type is belief-based discrimination, which arises from assumptions about an applicant's unobservable

characteristics inferred from an observable trait, such as gender. The second type is preference-based discrimination, which reflects a firm's inherent bias against certain races, sexes, or ages. Disentangling these forms of discrimination is crucial for understanding the dynamics at play.

From a regulatory perspective, identifying the different types of discrimination for an individual firm holds considerable importance. When discrimination stems from incorrect beliefs about an applicant, mitigating information asymmetries between the applicant and the firm can play a pivotal role in reducing discriminatory practices. On the other hand, if discrimination primarily arises from a firm's inherent bias against specific races, genders, or ages, regulators can then focus on targeting such firms for audits and investigations. This targeted approach ensures the enforcement of equal employment opportunities for all individuals.

To tackle the problem of identifying unit-level discrimination and the underlying source of this discrimination, we introduce a novel approach utilizing causal machine learning (ML) techniques. This method aims to generate unit-level discrimination estimates in situations where each unit within the studied group evaluates multiple individuals or items from a secondary group, and each individual or item in this secondary group is evaluated by multiple units. This framework applies to a broad spectrum of real-world interactions, including but not limited to applications for jobs, housing, credit, as well as online reviews.

In this chapter, we base our analysis solely on the outcomes of interactions between two entities. For instance, if we categorize the first entity as all firms and the second as all applicants, our required dataset must indicate whether an individual submitted an application to a firm. Furthermore, for those who applied, the data should reveal whether they received a callback from the firm.

Given the minimal data requirements, we first introduce a novel methodology for deriving the latent preferences of units and the characteristics of items or individuals from the outcomes observed. This method builds on the 'honest trees' concept from Athey and Imbens (2016) and incorporates it into collaborative filtering (CF) techniques. This method is notably effective for identifying sets of unobservable characteristics that are otherwise difficult to quantify. Additionally, it addresses the challenge of accurately capturing a firm's preferences, which can be problematic with conventional methods. Firms often harbor implicit preferences that are not readily apparent through direct measurement, underscoring the value of this innovative approach. In the second phase, we apply a Double Machine Learning estimator, developed by Chernozhukov et al. (2018), to obtain estimates of unit-level discrimination, while considering the influence of preferences and characteristics in both the matching and outcome processes.

To demonstrate our estimation approach, we simulate a gender-differentiated job application and hiring process within a hypothetical labor market. In this simulation, characteristics of male and female applicants, along with firms' preferences, are generated from unknown probability distributions. This setup allows us only to observe the submission of applications and the outcomes of interview selections. The simulation explores the influence of perceived match quality and potential gender-based discrimination on hiring decisions, employing a mechanism where interview invitations depend on surpassing a threshold influenced by match quality and gender biases, represented through a deterministic component and a Gaussian error term.

Furthermore, to test the real-world applicability of the proposed methodology, we collected a dataset of almost 150,000 reviews from Metacritic.com, encompassing over 8,000 films. The aim was to investigate whether film critics demonstrate gender-based discrimination or favoritism in their reviews of films directed by women. Using this methodology, the findings indicated that around 5% of critics showed a preference for films by female

directors. Alternatively, a ‘naive’ model that disregards critics’ preferences, the characteristics of the films, and the intrinsic review process suggested that more than 30% of critics gave more favorable reviews to films directed by women. This example emphasizes the importance of considering individual preferences in discrimination studies at the unit level, particularly in contexts where outcomes are significantly influenced by subjective judgments.

Over recent decades, the economics literature has extensively explored discrimination across various dimensions—gender, race, ethnicity, and religion—across multiple settings such as the labor market (Goldin, 1990), education (Alesina et al., 2018), law enforcement (Knowles et al., 2001), and marketplace (List, 2004). While this body of work largely examines discrimination at an aggregate level, focusing on average impacts, Becker (1957) seminal work on the economics of discrimination suggests that observed disparities at the aggregate level might result from a range of individual discriminatory behaviors. Our work contributes to this literature by providing evidence on unit-level discrimination.

However, our work is not the first in the literature that tries to address discrimination at an individual level. Ridgeway and MacDonald (2009) identified discriminatory behavior in a small fraction of New York City Police Officers in issuing pedestrian stops. Goncalves and Mello (2021) observed minority drivers receiving fewer ticket discounts in Florida, with more than 40% of officers showing bias. Vomfell and Stewart (2021) further examined police searches in the UK, finding widespread over-searching of ethnic minorities. Moreover, a series of studies investigated discrimination at an employment level for U.S. firms (Kline et al., 2023; Kline and Walters, 2021; Kline et al., 2022). This chapter aims to go a step further by separating biases from broader individual preferences, a distinction that has not been investigated in previous micro-level discrimination studies.

Our study introduces a novel approach for estimating discrimination at the individual level

with observational data, advancing beyond traditional methods that rely on comparing observed outcomes against potential outcomes based on real valuations (Knowles et al., 2001). While previous research often addresses observational study endogeneity through field experiments (Bertrand and Mullainathan, 2004), these experiments face limitations, such as biases from unobservable characteristics correlated with group identity (Heckman, 1998) the potential overemphasis on minor characteristics due to the forced similarity of candidates (Pager, 2007). Our method aims to overcome these challenges, providing a more nuanced analysis of discrimination.

The remainder of this chapter is organized as follows: The proposed methodology is described in detail in Section 3.2. Section 3.3 demonstrates the proposed estimation approach, and section 3.4 discusses the empirical exercise carried out. The chapter concludes with Section 3.5.

3.2 Conceptual Framework and Methodology

The conceptual framework of this study is based on the interaction between two distinct sets of entities: I and J . There is a many-to-many relationship between the items in the two sets in the sense that each item $i \in I$ is evaluated by one or more items $j \in J$, and similarly each item $j \in J$ evaluates one or many items in I . This setting is analogous to the application process in the job market where job seekers apply for potentially more than one employer, and employers receive and evaluate applications from potentially many applicants. Following this analogy, we refer to items $i \in I$ as applicants and items in J as employers throughout this chapter. The outcome of such interactions could be hiring decisions or callbacks for interviews.

Furthermore, each applicant is associated with a trait T (e.g., gender) that could be subject to bias from employers in the evaluation process. However, beyond potential

biases, employers might have preferences P over characteristics of applicants C that are known to be correlated with traits. The goal is to find out whether and which employers are likely to be biased in their evaluation processes beyond what is justifiable by the the employers' characteristics they desire.

The top panel in Figure 3.1 illustrates this process in a causal diagram. There are two routes through which traits T can influence the outcome Y : employers' biases displayed in the upper route, or through the *legitimate* lower path which capturing employers' desired characteristics among applicants. In the figure, the black nodes for T and Y denote the outcomes observable for the researcher. In the other hand, applicants' characteristics C and employers' biases θ and preferences P , which are not observed by the researcher, are displayed with the circles. The researcher is interested in estimating trait-based biased for one, many, or all entities in J . However,

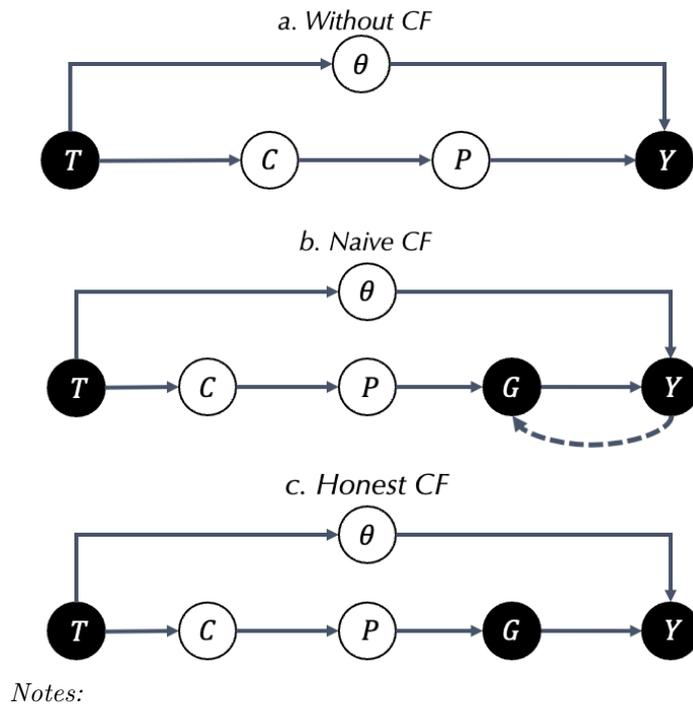


Figure 3.1: Conceptual Framework

Assume the researcher neither observes X nor g , and has no knowledge of the assignment

process M . All she observes are trait T , and outcomes $r_{i,j}$. Her objective is to estimate $\theta_j \in \Theta$ for every $j \in J$, as described by the equation:

$$r_{j,i} = \alpha + \theta_j T_i + g_j(X_i) + \varepsilon_{j,i} \quad (3.1)$$

At first glance, this task seems nearly impossible. How to factor in the unobserved attributes and diverse preferences when no direct evidence are available? The answer lies in a technique widely explored in machine learning: collaborative filtering (CF). This method assumes that if person A has the same opinion as person B on an issue, A is more likely to have B's opinion on a different issue than that of a random person. In essence, collaborative filtering creates a latent database of users' preferences, then uses this data to predict a user's tastes based on the tastes of similar users. This approach is widely used in various applications, such as recommending books, movies, or music, where the system suggests new items based on the likes and dislikes of similar users rather than analyzing the content of the items themselves.

Collaborative filtering (CF) encompasses a wide range of methods, each suited to various data availability scenarios. In this analysis, we have employed one of the fundamental methods, which requires minimal data input: Regularized Matrix Factorization (RMF), as described by Koren et al. (2009). This method starts with the matrix of reviews $R_{|I| \times |J|}$, where each entry $r_{i,j}$ indicates the assessment of reviewer j for individual i , and missing elements where no review was given. Typically, this matrix is *sparse*, as reviewers usually evaluate only a fraction of the total individuals. RMF effectively transforms the high-dimensional and sparse matrix R into two matrices of lower dimensions, $C_{|I| \times d}$ and $P_{|J| \times d}$, with $d \ll I, J$ being a hyperparameter of the model. The objective is for the matrix product $W_{|J| \times |I|} = CP'$ to effectively approximate the observed entries in R , while using regularization to avoid overfitting. For instance, using mean squared error loss and using

regularization on the sum of squared parameters in P and C , the loss function would be,

$$L = \sum_{(i,j) \in M} (r_{i,j} - p_j c_i)^2 + \left[\lambda \left(\frac{1}{|I|} \sum C^2 + \frac{1}{|J|} \sum P^2 \right) \right]$$

where M is the set of non missing entries, and λ is a regularization parameter.

The essence of Regularized Matrix Factorization (RMF) lies in its ability to distill the information in R into two condensed, lower-dimensional matrices: P and C . These matrices act as latent spaces, where P represents the latent preferences of the reviewers, and C captures the characteristics of the individuals being evaluated. In the process of constructing the latent spaces, the model positions similar reviewers and similar individuals/items close to one another in their corresponding matrices. This methodology effectively constructs a refined representation of the characteristics of both reviewers and reviewed individuals, based solely on the observed outcomes of their evaluations. RMF’s capability to infer rich characteristics from basic outcome data is what makes it a particularly powerful tool in recommendation systems.

However, applying standard CF methods in examining discrimination can be problematic, as it might incorporate micro-level biases into the resultant embeddings. To mitigate this, we propose a method similar to Athey and Imbens (2016)’s “honest trees”, termed “honest CF”. This approach begins by ensuring that trait-based biases are not incorporated into the latent preferences of reviewers. It involves factorizing the matrix $R_0 = [r_{i,j} | T_i = 0]$, which includes only the outcomes of individuals from the baseline group with trait $T = 0$. This approach effectively extracts reviewers’ latent preferences P^0 , uninfluenced by biases associated with the trait.

In the next step, the focus shifts to developing the latent space for individual characteristics, C . The central challenge here lies in separating the outcomes of a particular reviewer, from the representation of items that she has reviewed when estimating the reviewer’s

trait-based bias to avoid reverse causation. To achieve this, we employ a sample-splitting strategy, dividing the set of reviewers J into K distinct subsets J_1, J_2, \dots, J_K . Iteratively, we select one subset J_k and use the data from the remaining subsets $J - J_k$ to construct C^k . This procedure is repeated for each $k \in \{1, \dots, K\}$, leading to the creation of latent spaces for items that are constructed independently of the outcomes for a particular group of reviewers. Consequently, Equation 3.1 can be re-formulated in the following form,

$$r_{i,j} = \alpha + \theta_j T_i + g(P_j^0, C_i^k) + \varepsilon_{j,i} \quad (3.2)$$

There are two challenges in estimating the regression specified in Equation 3.2. First, the matching or assignment process has not been accounted for. Reviewers' preferences and individuals' characteristics are likely to affect the matching process between the two types of players as well as the outcomes of the evaluation. Second, function g needs to be estimated from the data. To overcome these challenges, we use the Double/Debiased ML (DML) method for treatment effect estimation as proposed by Chernozhukov et al. (2018).

To understand the idea behind DML, Consider the following partially linear regression framework as proposed in Robinson (1988):

$$\begin{aligned} Y &= \theta_0 D + g_0(X) + U, & \mathbb{E}[U|X, D] &= 0 \\ D &= m_0(X) + V, & \mathbb{E}[V|X] &= 0 \end{aligned}$$

In this framework, the first equation models the relationship between the treatment variable D and the outcome variable Y . X denotes the vector of control variables influencing both the assignment to treatment and outcomes through unknown functions $m_0(X)$ and $g_0(x)$. The terms V and U represent disturbance variables. This model assumes that con-

ditional on observable features X , the treatment assignment is effectively random, i.e., D is conditionally exogenous. Consequently, θ_0 can be interpreted as the causal effect of treatment on the outcome.

After minor adjustments, the DML framework can be adapted to estimate micro-level discrimination. This is represented by the model:

$$\begin{aligned} r_{i,j} &= \theta_j T_{i,j} + g(P_j^0, C_i^k) + \varepsilon_{j,i} \\ T_{i,j} &= m(P_j^0, C_i^k) + \epsilon_{j,i} \end{aligned} \tag{3.3}$$

While the first equation of the model mirrors the interpretation of the earlier framework, the addition of the second equation captures the influence of preferences and characteristics on the matching process between individuals and reviewers.

Algorithm 2 summarizes the proposed methodology to obtain micro-level estimates of discrimination.

Algorithm 2 Estimating Micro-Level Coefficients of Discrimination

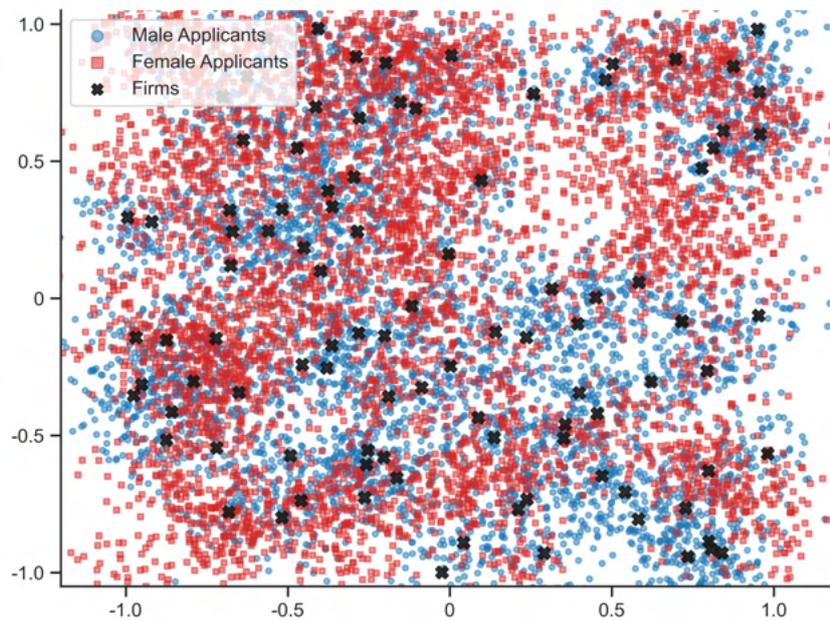
1. Factorize the matrix of outcomes R into P^0 and C^0 .
 2. Split the set of reviewers J into K mutually exclusive subsets $J = \{J_1, \dots, J_K\}$.
 3. for J_k in $\{J_1, \dots, J_K\}$:
 - 3.1 Construct the critic-movie ratings matrix R_{-k} for all critics not in J_k ;
 - 3.2 Factorize R_{-l} to obtain C^k ;
 4. Obtain DML estimates of the model specified in Equation 3.3.
-

3.3 Simulated Example

This section describes a simulated exercise aimed at illustrating the estimation procedure in a hypothetical labor market scenario. The exercise simulates a labor market, wherein job seekers of two genders apply for open positions, and employers decide whether to interview them. Both the job seekers and the hiring firms' characteristics are generated

from probability distributions that remain unknown to the econometrician. The econometrician only observes the applications sent by the job seekers and the subsequent interview decisions made by the firms.

In this exercise, we model the characteristics of male and female applicants as being separately drawn from N_C randomly generated clusters within a \mathcal{D}_A -dimensional isotropic Gaussian distribution, each with a standard deviation of σ . The centroids of these clusters are uniformly selected from intervals ranging between $(-1, +1)$. Despite male and female characteristics originating from distinct clusters, the random positioning of cluster centroids results in regions where these characteristics overlap. In addition, there are N_E employers, and their preferences are drawn from a uniform distribution over a \mathcal{D}_F -dimensional space, bounded between -1 and +1. For the sake of simplicity, let's assume $\mathcal{D}_A = \mathcal{D}_F = 10$. Figure 3.2 plots the positioning of applicants and firms within the first two dimensions of their respective latent characteristic and preference spaces.



Notes: The figure plots the first two dimensions of applicants' characteristics and employer preferences latent spaces.

Figure 3.2: Simulated Latent Characteristics and Preferences

To model the job application process, we introduce a simple random mechanism. Al-

though sending job applications incurs no cost, exogenous factors, such as timing and geographical constraints, naturally limit the pool of employers to which applicants can feasibly apply. Consequently, each applicant ends up applying to only a subset of all potential employers. Specifically, we simulate each applicant applying randomly to n_i employers, where n_i is drawn from a Poisson distribution with parameter λN_F . Here, λ represents the expected fraction of employers that receive applications from each applicant. Employer j 's assessment of applicant i is swayed by two pivotal factors: the perceived match quality and potential gender-based discrimination. An invitation for an interview is sent to the applicant if the firm's evaluation surpasses a predetermined threshold, ω . Thus, the evaluation process can be represented as follows:

$$y_{i,j} : \begin{cases} 1 & \text{if } g(c_i, p_j) + T_i \times \theta_j + e_{i,j} > \omega \\ 0 & \text{otherwise} \end{cases}$$

Here, $g(c_i, p_j)$ denotes the deterministic component of the match quality between the applicant and the employer, and $e_{i,j}$ is a Gaussian error term. We assume $g(c_i, p_j)$ to be the cosine similarity between the applicant's characteristics vector c_i and the firm's preference vector p_j . Lastly, θ_j captures the gender-based biases, assumed to be drawn from the standard normal distribution.

The econometrician is presented with an outcome matrix similar to the following example,

$$Y_{N_A \times N_E} = \begin{bmatrix} 1 & - & 1 & - & \cdots & 0 \\ 0 & - & 0 & - & \cdots & - \\ - & 0 & - & 1 & \cdots & - \\ 0 & - & - & - & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & - & 0 & - & \cdots & 1 \end{bmatrix}$$

This matrix reveals three possible outcomes for each applicant-firm pair: either the applicant has not submitted an application to the firm, the applicant has applied but has not received an interview invitation, or the case where the applicant has applied and has been called back for an interview.

Table 3.1 summarizes the parameters and the modeling assumptions used in the simulation.

Description	Notation	Value/Func. Form
Number of employers	N_E	100
Number of applicants, men	N_A^M	5000
Number of applicants, women	N_A^W	5000
Employers' references space	P	$U(-1, 1)^{10}$
Applicants' characteristic space	C	Gaussian clusters in \mathbb{R}^{10}
Number of skill cluster centroids	-	100
Skill clusters std	-	0.1
Number of applications per applicant	-	$Pois(0.2N_E)$
Interview Threshold	ω	1
Perceived Match value	-	$g(c_i, p_j) + \theta_j + \varepsilon_{i,j}$
Deterministic match value function	g	$5 \cos(c_i, p_j)$
Employers' gender discrimination	θ	$N(0, 1)$
IID Random Noise	ε	$N(0, 1)$

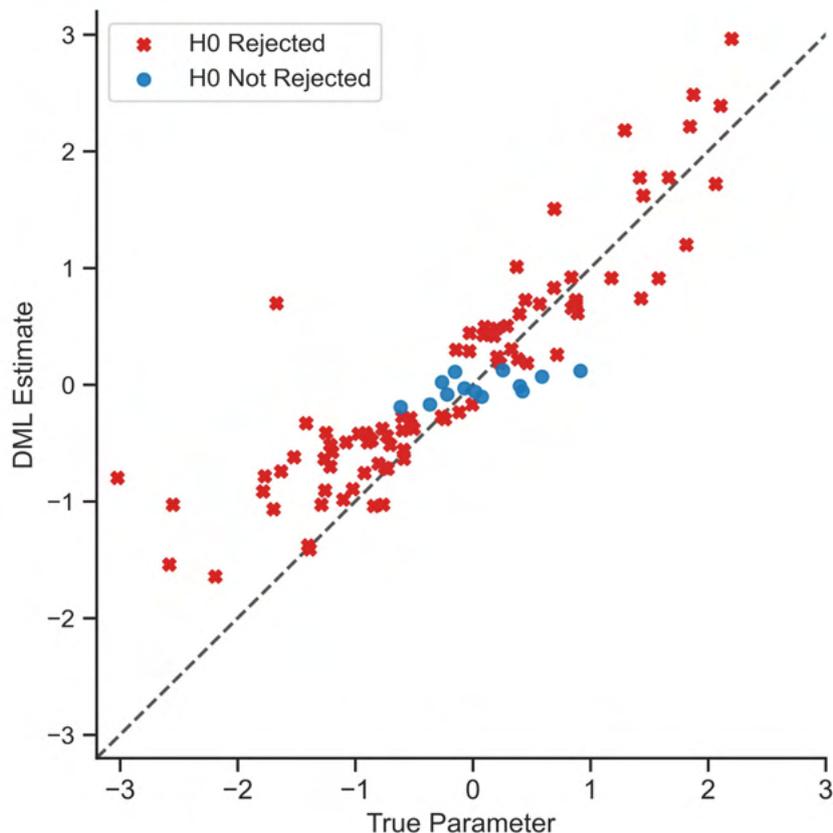
Table 3.1: Baseline Choices of Parameters

After observing the outcomes Y , we apply the procedure described in Algorithm 2 to obtain the estimates of gender biases of the employers. In the first step, we must estimate P^0 , employers' preferences, using only the outcomes of the baseline group, i.e., men with $T = 0$. The second step involves splitting employers into 10 mutually exclusive subsets and obtaining 'honest' estimates of applicants' characteristics in one subset, using the outcomes in the other subsets¹. To test the robustness of our approach to misspecification of the preferences and characteristics spaces, we estimate the latent spaces by applying RMF with $d = 8$, where the actual spaces are in \mathbb{R}^{10} .

¹There is a negligible probability that certain applicants have applied exclusively to firms within a single subset. For these applicants, characteristics cannot be estimated, and their observations are excluded from subsequent analysis.

Lastly, the DML regression model specified in Equation 3.3 is estimated. For the treatment model, we use a logistic classifier with $L2$ regularization. For the outcome model, we employ a Random Forest model.

Figure 3.3 plots the estimated coefficients of employers' bias versus the true parameter value. Despite the misspecification of latent preference spaces and the match value function, the estimates are still closely centered around the 45-degree line. To assess the statistical significance of the estimates, the Benjamini and Hochberg (1995) method for multiple hypothesis testing with a false discovery rate set to 0.05 is employed. The null hypothesis of $\theta = 0$ is rejected in only four out of 100 instances when the estimated coefficient exhibits the wrong sign.



Notes:

Figure 3.3: Estimates of Employers Bias

3.4 Empirical Example

The ensuing section presents an empirical application of the established methodology using a real-world dataset. This approach is instrumental in demonstrating the framework’s efficacy to discern micro-level discrimination in practical settings.

3.4.1 Data Description

A dataset comprising film reviews from professional critics was constructed using data from Metacritic.com, a review aggregator website. Metacritic collects reviews from approximately 100 sources, assigning ratings on a uniform scale of 0-100. These ratings were transformed to a 0-1 scale for this analysis². The objective is to apply the methodology described in Section 3.2 to explore potential discriminatory patterns in critics’ reviews of movies directed by women.

Metacritic provides detailed information, such as the names of film directors. The gender of directors was deduced using their first names and the *gender-guesser* Python library³, a prevalent tool for name-based gender inference. Entries were removed if *gender-guesser* was unable to make a prediction (name not found in its database) or if the name was non-specific to a particular gender. This gender identification procedure was verified for accuracy against a Wikipedia directory of female directors⁴, with a misclassification rate under 5%. To maintain simplicity in the analysis, films with more than one director were excluded, as over 95% of movies in the dataset had a single director.

Table 3.2 offers a summary of statistics for selected variables in the dataset. Data was collected for films released from 1990 to 2021 and having at least seven critic reviews on

²In cases where an explicit rating is absent, Metacritic’s evaluators assign a score reflecting their assessment of the article.

³<https://pypi.org/project/gender-guesser/>

⁴https://en.wikipedia.org/wiki/List_of_female_film_and_television_directors

	Count	Mean	Std.	Min	Med	Max
Year	8,284	2008.7	8.07	1990	2010	2021
Critic Rating	145,522	0.631	0.210	0	67	100
Films' N.o. Critic Reviews	8,284	17.6	9.00	1	16	47

Notes: The table shows summary statistics for the selected variables in the data. The data is limited to the reviews from critics who have evaluated at least 30 movies directed by women.

Table 3.2: Summary Statistics of Selected Variables

Metacritic. This was further narrowed down to critics who had reviewed a minimum of 30 films directed by female directors. The filtered dataset contains over 145,000 reviews from 205 critics, spanning around 8,300 films and 3,900 directors. Films directed by women constitute nearly 14% of the dataset. Each film, on average, garnered reviews from more than 17 critics, with an average rating of 0.63 on a 0-1 scale.

3.4.2 Estimation

As underscored earlier in this study, the estimation of discrimination or favoritism at the individual level is of considerable significance for several reasons. Chief among them is the potential for aggregate-level bias estimates to be misleading. To illustrate, consider a hypothetical scenario in our context: a seemingly minor bias against movies directed by women could arise either from a general absence of discrimination among critics or from the presence of two distinct groups of reviewers – one disproportionately critical and the other overly favorable towards female-directed films. While both scenarios lead to similar estimates of aggregate-level bias, they depict starkly different realities of micro-level discrimination.

In the assessment of individual-level biases or favoritism, the role of personal preferences among decision-makers is pivotal. For instance, in this analysis, a critic's preference for particular genres or themes – more frequently found in films directed by either gender – might inadvertently color their reviews. This genre or theme preference could manifest as

apparent gender bias in reviews, while it truly stems from the critic’s own cinematic tastes. Overlooking these personal preferences risks incorrectly categorizing critics as biased.

Therefore, we implement the method outlined in Section 3.2 to estimate individual-level bias/favoritism regarding critics’ evaluation of female-directed films. The approach involves estimating the following Double Machine Learning (DML) model

$$\begin{aligned} r_{i,j} &= \theta_j FD_{i,j} + g(P_j^0, C_i^k) + \varepsilon_{j,i} \\ FD_{i,j} &= m(P_j^0, C_i^k) + \epsilon_{i,j} \end{aligned} \tag{3.4}$$

Here, $r_{i,j}$ represents the rating given by critic j to film i , while $FD_{i,j}$ is a binary indicator denoting whether film i was directed by a woman. The parameter θ_j is indicative of the critic-specific bias/favoritism towards films directed by women.

In this dataset, typically, critics review only a limited selection of films, and correspondingly, each film is assessed by a small group of critics. With a total of over 8,000 films and approximately 200 critics, the dataset comprises less than 150,000 observed ratings, indicating that under 10% of all possible ratings are recorded. Notably, the use of Collaborative Filtering in industrial settings is intended to predict ratings that a user might assign to items they have not yet reviewed (such as books, music, or movies) and to recommend items likely to be highly rated by the user.

As outlined in Algorithm 2, the process begins with applying regularized matrix factorization (RMF) to decompose the observed rating matrix $R_{I \times J}$ into $C_{I \times d}^0$ and $P_{J \times d}^0$ and . RMF starts by randomly initializing matrices P and C , followed by employing optimization techniques such as gradient descent or its variants to minimize the regularized loss function, detailed in Section 3.2. This step involves selecting the embedding dimension d and the regularization parameter λ . The value of d was fixed at 100, a commonly adopted figure. For determining λ , a trial and error method was employed, using 10% of the data

as a test set and testing several multiples of 10 as potential values for λ . This process led to the selection of $\lambda = 0.01$, which produced a mean squared error (MSE) of 0.074 in the test set.

In the subsequent phase, RMF was implemented following the initial stage outlined in Algorithm 2. Specifically, the RMF algorithm was applied exclusively to the matrix of ratings for films directed by men, omitting the use of a test set. This application aimed to generate P^0 , signifying the matrix of critics' latent preferences, deliberately isolated from their evaluations of films directed by women.

Upon deriving P^0 , the second step of Algorithm 2 involved randomly dividing critics into $K = 10$ subsets. For each subset k , RMF was then applied to the ratings matrix R_{-k} , comprising ratings data from critics in the remaining subsets, to generate C^k . Here, C^k indicates the film characteristics' embeddings, isolated from the ratings by critics in that particular subset.

In the final step, P^0 and the C^k matrices were utilized to derive DML estimates for the model specified in Equation 3.4. For these estimates, a binary logistic classifier with L2 regularization was employed as the learner for m in the equation. The regularization parameter was set to the default value of 1, as specified in the *scikit-learn* package. For the learner g , a Random Forest regression was chosen, using the hyperparameters outlined in the *DoubleML* package documentation for the DML estimator in partially linear regression models^{5 6}.

To draw a comparison between the outcomes derived by the proposed method and those from a conventional approach, we also estimated the following Ordinary Least Squares

⁵<https://docs.doubleml.org/stable/api/generated/doubleml.DoubleMLPLR.html>

⁶Since my goal here is to clarify the proposed methodology, we did not engage with hyper-parameter tuning or comparing the results using alternative learning models. However, in practice, trying with different models, and hyper-parameter tuning will be generally helpful to examine the overall robustness of the results.

(OLS) model:

$$r_{i,j} = \alpha + \beta_j FD_{i,j} + \gamma_j + e_{j,i}$$

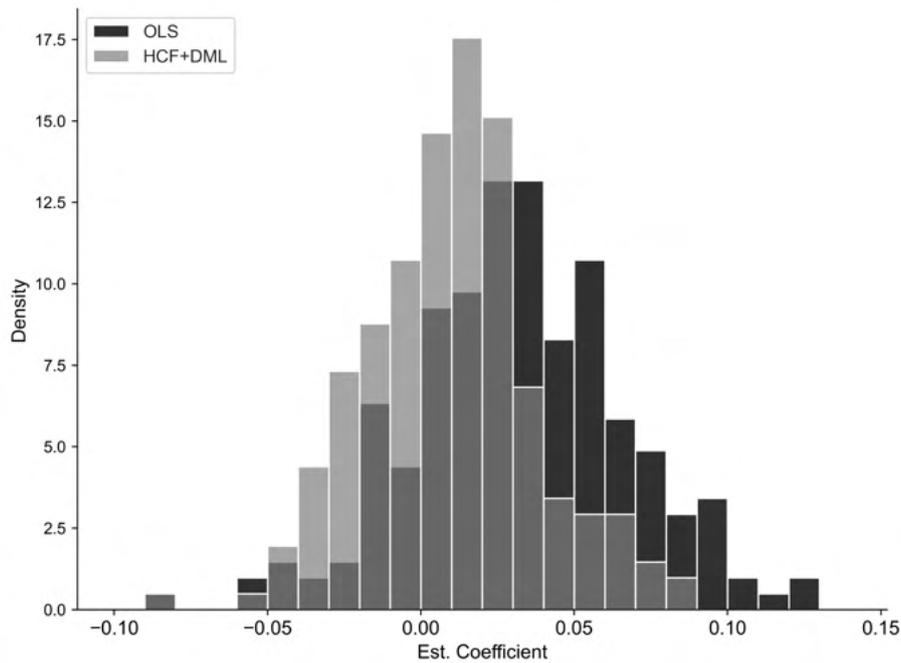
In this model, β_j represents the OLS estimate of critic j 's discrimination/favoritism towards films directed by women. The terms γ_j denotes the critics' fixed effects .

Figure 3.4 illustrates the distribution of the estimated θ values from Equation 3.4 using the proposed method, denoted as 'HCF+DML', and compares it with the distribution of the estimated β values from Equation 3.4.2, labeled 'OLS'. Although there is an overlap between the two distributions, notable differences are evident. The OLS estimates are centered around 0.04, approximately. In contrast, the DML estimates seem to be generally smaller, and centered around, 0.01. Overall, while the mean differences in the two distributions (i.e., the aggregated estimate of discrimination/favoritism) are minor (around 0.03 scores), there is a marked difference in the distribution tails..

The findings corroborate the hypothesis that aggregate-level estimates might not truly represent the actual distribution of discriminatory behaviors at the individual level. Figure 3.4 shows that the distribution of OLS estimates, with its heavier right tail on the positive side, implies a greater likelihood of identifying critics who positively discriminate for films directed by women. However, the results from the method employed in this study paint a different picture. Once preferences and characteristics are accounted for, the distribution of estimates indicates far smaller share of critics with high levels of discrimination based on films' directors genders.

To identify critics whose estimates of discrimination or favoritism are statistically significant from zero, we employed the method by Benjamini and Hochberg (1995) for multiple hypothesis testing⁷. Figure 3.5 contrasts the estimated coefficients from the OLS model

⁷Given the large number of tests, using standard confidence intervals to reject the null hypothesis is not appropriate. Considering a p-value threshold of less than 0.05 for rejecting the null hypothesis of $\beta_c = 0$ in the context of testing 100 estimates would lead to approximately five rejections due to random variation alone. The method by Benjamini and Hochberg (1995) addresses this by controlling the false



Notes: The figure displays the distribution of the micro-level estimates of discrimination obtained via the proposed methodology (HCF+DML) outlined in Equation 3.4, and the OLS estimates of the model presented in Equation 3.4.2.

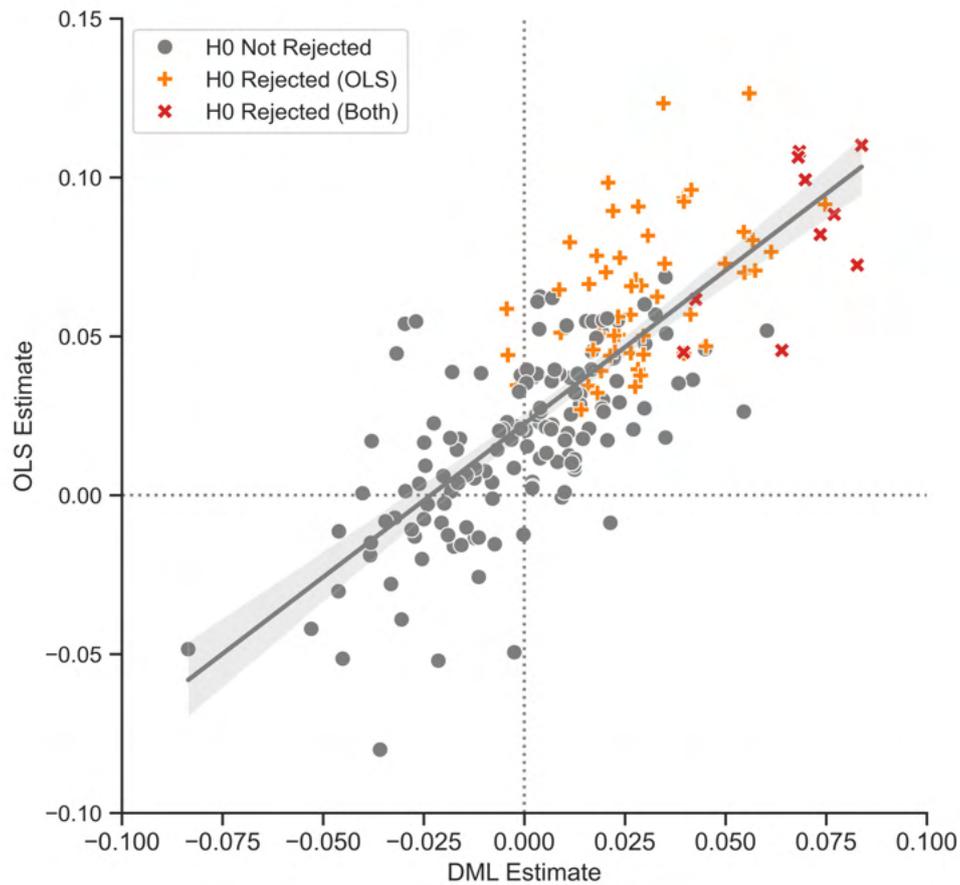
Figure 3.4: Distribution of Estimated Coefficients

with those obtained via HCF+DML. Although there is a strong correlation between the two sets of estimates, notable differences emerge when assessing the coefficients statistically different from zero using the Benjamini and Hochberg (1995) method with a false discovery rate of 0.10. In the OLS, the null hypothesis is rejected for 64 estimates (all positive). In contrast, the DML estimates reveal 10 statistically significant coefficients.

The findings of this analysis reveal that while factoring in preferences, characteristics, and the matching method as per the proposed methodology in this chapter might marginally adjust the overall estimate of discrimination, it significantly alters the micro-level distribution of these estimates. A naive approach, which omits these aspects, suggested that about 30% of critics demonstrate favoritism or discrimination, based on directors' gender. In contrast, the approach used in this study, accounting for these three essential factors,

discovery rate (FDR), i.e., the probability of incorrectly rejecting a true null hypothesis.

showed that only about 5% of critics show patterns of favoritism toward films directed by women beyond what their preferences and the films' characteristics would justify.



Notes: The figure plots the OLS estimates versus the DML estimates of individual-level discrimination among critics. The statistical significance is tested using Benjamini-Hochberg method using a false discovery rate of 0.10.

Figure 3.5: OLS versus DML Estimates

3.5 Conclusion

In this study, we proposed a novel methodology to obtain micro-level estimates of discrimination in a reviewer-applicant setting, that accounts for unobservable preferences, characteristics, and a potentially endogenous matching process. The study proposes ‘Honest’ Collaborative Filtering, a method to extract latent preferences and characteristics partly isolated from the observed behavior of the individuals. As an empirical example demon-

strating the method's use-cases in practice, we analyzed the performance of this method using real-world data from film critic reviews to test for critics' discrimination/favoritism based on gender of the directors. The results suggests that while the aggregate-level estimate of discrimination/favoritism obtained using the proposed method are close to the ones obtained via a naive approach that disregards preferences and characteristics on the two sides, the micro-level estimates provides considerably different pictures on the underlying distribution of discrimination/favoritism.

Bibliography

- Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2018). Revealing stereotypes: Evidence from immigrants in schools. Technical report, National Bureau of Economic Research.
- Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27), 7353–7360.
- Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.
- Benjamini, Y. and Y. Hochberg (1995, January). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1), 289–300.
- Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. *Handbook of economic field experiments* 1, 309–393.
- Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review* 94(4), 991–1013.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, February). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1), C1–C68.
- Goldin, C. (1990). *Understanding the gender gap: An economic history of American women*. Oxford University Press.
- Goncalves, F. and S. Mello (2021). A few bad apples? racial bias in policing. *American Economic Review* 111(5), 1406–1441.
- Heckman, J. J. (1998). Detecting discrimination. *Journal of economic perspectives* 12(2), 101–116.

- Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics* 137(4), 1963–2036.
- Kline, P., E. K. Rose, and C. R. Walters (2023). A discrimination report card. *arXiv preprint arXiv:2306.13005*.
- Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica* 89(2), 765–792.
- Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of political economy* 109(1), 203–229.
- Koren, Y., R. Bell, and C. Volinsky (2009, August). Matrix Factorization Techniques for Recommender Systems. *Computer* 42(8), 30–37.
- List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics* 119(1), 49–89.
- Pager, D. (2007). The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science*, 104–133.
- Ridgeway, G. and J. M. MacDonald (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association* 104(486), 661–668.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.
- Vomfell, L. and N. Stewart (2021). Officer bias, over-patrolling and ethnic disparities in stop and search. *Nature Human Behaviour* 5(5), 566–575.