The Economics of Regulation for the AI Supply Chain

Sihan Qian
School of Economics and Management, Tsinghua University, qsh22@mails.tsinghua.edu.cn

Amit Mehra
Naveen Jindal School of Management, The University of Texas at Dallas, amit.mehra@utdallas.edu

Dengpan Liu
School of Economics and Management, Tsinghua University, liudp@sem.tsinghua.edu.cn

Abstract. The rise of foundation models has led to the emergence of the AI supply chain, where upstream foundation model providers offer fine-tuning and inference services to downstream firms developing domain-specific applications. During fine-tuning, downstream firms pay providers to leverage their computing infrastructure to fine-tune models using proprietary data, creating a co-creation dynamic that enhances these models' quality. Given growing concerns that foundation model providers and downstream firms might capture too much consumer surplus, along with increasing global policy measures to regulate the AI supply chain, this study employs a game-theoretic model involving one foundation model provider and two competing downstream firms to analyze how various policy interventions impact consumer surplus within this ecosystem. Our analysis shows that policies promoting price competition in downstream markets (i.e., *pro–price–competitive policies*) boost consumer surplus only when compute or data preprocessing costs are relatively high. In contrast, compute subsidies are effective only when these costs are relatively low, suggesting that the two policy approaches can complement each other. Policies promoting quality competition in downstream markets (i.e., *pro–quality–competitive policies*), on the other hand, consistently improve consumer surplus, regardless of cost levels. We also find that under pro–price–competitive policies or compute subsidies, both the foundation model provider and downstream firms can achieve higher profits alongside greater consumer surplus, leading to a win–win–win outcome. However, pro–quality–competitive policies increase the provider's profits while consistently reducing the profits of downstream firms. Lastly, as compute costs decline, both the foundation model provider and consumers benefit, though downstream firms may see reduced profitability, highlighting the need for carefully calibrated policy interventions. Declining compute costs also make pro–price-competitive policies less effective, while increasing the effectiveness of compute subsidies in promoting consumer surplus. These findings offer important strategic insights for foundation model providers, downstream firms, and policymakers seeking to foster an AI supply chain that is both economically efficient and socially beneficial.

Key words: Artificial intelligence, Foundation model, Fine-tuning, Data, Competition

1. Introduction

The rapid advancement of artificial intelligence (AI) is transforming economic sectors on an unprecedented scale (Sun et al. 2025). The global AI market, valued at $538.13 billion in 2023, is projected to grow at a compound annual growth rate (CAGR) of 19.1% to reach $3,680.47 billion by 2034 (Precedence Research 2023). However, this profound growth introduces significant risks to consumer welfare (FTC 2025b). This dual nature of AI, functioning as both a powerful productivity tool and a source of risk, creates an urgent need

for effective regulatory oversight. However, applying traditional regulatory frameworks to AI is problematic because the technology operates within a complex, multi-layered market structure, often conceptualized as an *AI supply chain* (Cen et al. 2024a). This structure is characterized by a distinct upstream-downstream dynamic: a small number of upstream providers develop large-scale AI models, also known as *foundation models*, which are then adapted and deployed by a broad downstream market of firms for specific applications. The inherent interdependence of this structure renders existing regulatory frameworks inadequate, inevitably leading to substantial gaps in oversight (Cen et al. 2024b). Therefore, protecting consumers and ensuring market efficiency requires a holistic regulatory approach that is specifically designed to address the unique interdependencies of the AI supply chain.

1.1. Foundation Models

To develop a holistic regulatory approach to the AI supply chain, it is essential to first gain a comprehensive understanding of its most fundamental component: the upstream segment, where developers create foundation models. These models, built with billions of parameters and trained on massive datasets, exhibit remarkable capabilities in tasks such as logical reasoning and natural language processing (Vohra 2024). Often hailed as a cornerstone of the next industrial revolution (Menear 2024), foundation models have driven major technology companies, including Microsoft, Amazon, and Google, to make substantial investments to secure dominance in the upstream market, with Google alone reportedly committing over $100 billion (Bhati 2024).

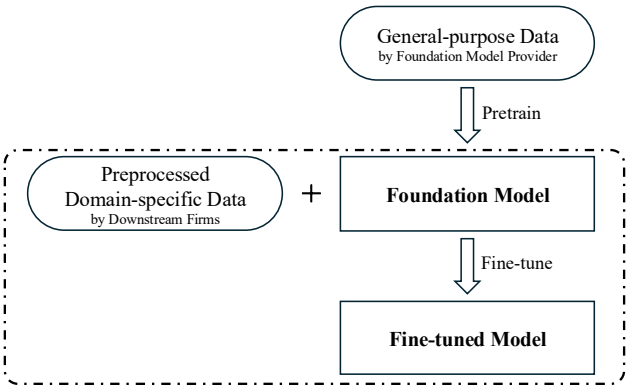
Despite their impressive general-purpose performance, foundation models often fall short in specialized domains such as healthcare and law. These limitations stem primarily from restricted access to high-quality, domain-specific data, which is often subject to privacy regulations and legal requirements (Sala 2024). Organizations in these sectors generally lack the resources to develop foundation models independently due to the prohibitively high costs involved, which include not only massive computational resources but also specialized technical expertise. For example, training a model on the scale of GPT-4 is estimated to cost over $100 million, making in-house development infeasible for most organizations (Wikipedia 2023).

1.2. Fine-tuning

To tackle these challenges, a co-creation process has taken shape within a comprehensive AI ecosystem, bringing together upstream foundation model providers and downstream firms. As shown in Figure 1, adapting a model for downstream applications typically requires two key training stages. The first stage is pretraining, which involves building a foundation model using general-purpose data (Stryker 2025). Pretraining is primarily conducted by major technology companies such as Google and OpenAI (Singh 2024).

The second stage, known as fine-tuning, is typically carried out by downstream firms and involves two critical steps. The first step is preprocessing domain-specific data to remove noise, errors, and inconsistencies[1](Huizenga 2025). Once the data is cleaned, it is used to retrain the foundation model, resulting in a

Figure 1 Creation of a fine-tuned model



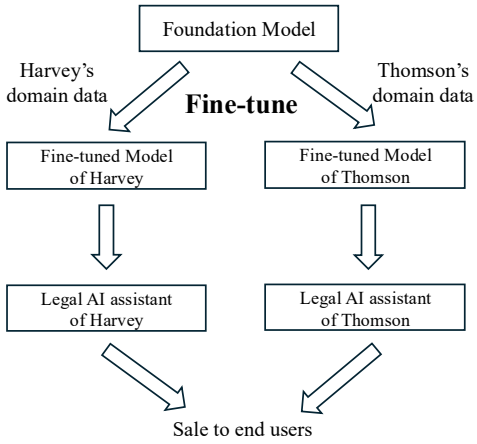
fine-tuned version tailored to specific downstream applications. Unlike the resource-intensive pretraining phrase, fine-tuning requires significantly less computational power, making it a feasible option for most firms (Competition and Markets Authority 2023). This reduced computational cost allows smaller companies to leverage models developed by large tech firms, adapting them to their own specialized needs (AI Now Institute 2023). Fine-tuning is widely regarded as essential for embedding complex, domain-specific behavioral patterns into AI models (OpenAI 2023). Supporting this, a 2024 report by PricewaterhouseCoopers (PwC) finds that organizations using fine-tuned models experienced a 37% increase in the accuracy of AI-generated content. The same report predicts that by 2030, 90% of companies will have deployed at least one fine-tuned large language model (ITSoli 2025).

To protect the proprietary nature of the foundation model's parameters, model providers do not share the models directly with downstream firms. Instead, firms are required to upload clean, domain-specific data, obtained by preprocessing their datasets, incurring what are known as *data preprocessing costs*, to the provider's infrastructure for fine-tuning. Given that fine-tuning demands significant computational resources, providers also face *compute costs*. In exchange, they charge a *fine-tuning fee*, typically based on the volume of preprocessed data firms use on the provider's infrastructure. Once the model has been fine-tuned, it is hosted on the provider's infrastructure, and firms are charged an *inference fee*, usually based on the volume of usage data (both input and output), to access the model for downstream applications (OpenAI 2024b). In this way, foundation model providers generate two primary revenue streams: one from fine-tuning fees and the other from inference fees.

1.3. The AI Supply Chain

The central role of fine-tuning in AI development has led to the emergence of the AI Supply Chain (Cen et al. 2024a), where foundation model providers offer general-purpose models that downstream firms fine-tune for specialized applications and subsequently market to end users. A real-world example of this AI supply chain is illustrated in Figure 2. For instance, law firms like Harvey and Thomson Reuters are leading players in developing AI-powered legal assistants (Fisher 2024). Both firms fine-tune OpenAI foundation models

Figure 2 Example of an AI supply Chain



using their proprietary legal data (OpenAI 2024a, Reuters 2024). Leveraging these fine-tuned models, they have created and commercialized advanced AI legal assistants that help attorneys with tasks such as drafting legal documents and managing complex litigation (OpenAI 2024a).

This novel form of supply chain has several distinctive characteristics. First, the foundation models are fine-tuned through a co-creation process involving both the foundation model provider and the downstream firm. This approach to product development differs fundamentally from traditional supply chains, where manufacturers typically bear sole responsibility for product design and development. Even in collaborative supply chains—such as those involving vendor–client partnerships for enterprise systems like ERP (Demirezen et al. 2020, Gupta et al. 2023)—the co-creation process generally occurs independently across vendors and clients. In contrast, the fine-tuning of foundation models entails a mutually dependent form of co-creation: upstream providers supply not only the base foundation models but also the computational infrastructure necessary for fine-tuning, while downstream firms contribute domain-specific data and expertise to adapt these models for specialized applications. Second, the provision of computational resources for fine-tuning creates a revenue stream for foundation model providers during the co-creation process, which is a unique form of revenue generation in contrast to traditional settings, where manufacturers generate revenue only through wholesale prices. Third, the fine-tuned models represent customized versions of foundation models that are deployed through a Software-as-a-Service (SaaS) framework. Traditionally, software customization occurs in on-premise settings, where clients adapt the software locally; in such cases, SaaS-style pricing models are typically not applied. Hence, the integration of extensive customization with a SaaS deployment model constitutes a distinctive feature of our setting.

1.4. Motivation and Research Questions

The rapid diffusion of AI-enabled products has introduced substantial risks to consumer welfare, particularly by shifting power from consumers to firms. As Nobel laureate Daron Acemoglu cautions, the use of AI technologies and data increases the power of companies over consumers, with important distributional

implications that enable AI-intensive companies to capture a larger share of consumer surplus (Acemoglu 2021, p. 13).

These concerns are compounded by the fact that AI-enabled products introduce fundamentally new types of services for consumers (e.g., AI-based legal assistants), yet assessing their quality remains inherently difficult (Ngo 2025). For example, consumers cannot easily anticipate the extent of hallucinations, which can significantly undermines the reliability of outputs. Legal experts caution that AI systems remain susceptible to errors, biases, and malfunctions, requiring professionals to be mindful of their limitations (Eldridge 2025). More troubling, some companies have exploited AI hype, misrepresenting their products' capabilities and further complicating consumers' ability to evaluate quality. Take, for example, DoNotPay, which marketed itself as the "world's first robot lawyer," falsely claiming to replace human legal expertise but ultimately failing to meet those promises (FTC 2025b). This lack of quality transparency allows downstream firms to soften competition, potentially leading to low-quality offerings and a decline in consumer welfare. In a similar vein, challenges also arise in pricing: Opaque and highly differentiated pricing schemes hinder straightforward comparisons. As Dreyer (2025) notes, reduced comparability undermines competitive pressure and tends to drive prices upward.

Given these issues, a natural regulatory response to AI supply chains is to implement pro-competitive policies that strengthen price or quality competition in downstream markets(Austrian Federal Competition Authority 2024). Such approaches have already been successfully applied in the broader digital economy. To promote price competition, for example, the U.S. Federal Trade Commission (FTC) required Ticketmaster to display the full price of tickets upfront during the purchase process, thus improving price transparency (Boggs 2025). In another case focusing on quality competition, the FTC alleged that Fashion Nova, a U.S.-based online fast-fashion retailer, misled consumers by suppressing negative reviews, underscoring the importance of truthful product quality disclosure (FTC 2025c). Some of these established regulatory tools have already been adapted for AI-enabled product markets. For example, the FTC has begun requiring firms to refrain from making deceptive claims about the capabilities of consumer-facing AI products (Hrdy et al. 2025, Leblanc et al. 2025). Such measures contribute to a broader competition policy framework by curbing misleading practices that distort consumer choice and weaken competitive pressures (FTC 2025).

Another regulatory strategy targets upstream foundation model providers with the AI supply chain, aiming to enhance consumer surplus by subsidizing the fine-tuning process and incentivizing the development of higher-quality products and services in downstream markets. A notable example is the annual commitment of the Beijing Municipal Government of 100 million RMB in computing power subsidies to firms engaged in large-scale AI model services (Wu 2024). Similar programs have been introduced by municipal governments in Shenzhen, Shanghai, Chengdu, and several other Chinese cities. By partially offseting the substantial costs of compute for fine-tuning, these subsidies help lower barriers to producing higher-quality AI-enabled offerings.

The effectiveness of regulatory interventions in the AI supply chain, however, is likely to differ from outcomes seen in traditional supply chain models, due to the unique co-creation dynamics inherent in this intricate ecosystem. In conventional supply chains, pro-competitive policies that lower retail prices typically enhance consumer surplus without diminishing the manufacturer's incentive to maintain product quality. In contrast, in the AI supply chain context, a reduction in downstream prices can erode the marginal returns of downstream firms, weakening their motivation to invest in fine-tuning and thereby diminishing overall product quality. This interdependence between pricing, incentives, and quality gives rise to a welfare trade-off absent in conventional market structures.

This trade-off highlights the need to evaluate not only whether regulatory interventions enhance consumer welfare, but also the mechanisms through which they operate and the conditions under which they may inadvertently discourage quality investment. Understanding these mechanisms is essential because the welfare implication of a policy depends on how it shapes firms' incentives: Measures that appear pro-competitive may, under certain conditions, weaken firms' motivation to invest in quality, reducing overall welfare. Recent debates on AI regulation reveal growing gaps in adapting existing regulatory frameworks to complex AI supply chains, underscoring the importance of understanding how different policy instruments perform in such environments (Cen et al. 2024b). These considerations motivate our central research questions: *How do regulatory approaches that promote competition in downstream markets or provide compute subsidies impact consumer surplus within AI supply chain, and which policy instrument proves most effective under varying conditions?* Addressing these questions is crucial to ensure that regulatory efforts ultimately promote, rather than compromise, welfare in AI-enabled markets.

The success of any regulatory policy ultimately depends on its acceptance by industry participants. Therefore, it is crucial to understand how pro-competitive policies and compute subsidies impact the profitability of both downstream firms and foundation model providers. Policies that increase consumer surplus while simultaneously improving the profitability of providers and downstream firms are more likely to be adopted, as it creates a "win–win–win" outcome in which all stakeholders benefit. Conventional wisdom suggests that heightened competition erodes the market power of downstream firms, thereby reducing their profitability (Matsushima 2006). However, Tyagi (1999) provides an alternative view, showing that increased downstream rivalry can induce price adjustments by upstream providers, which may ease competitive pressures among downstream firms and mitigate the need for aggressive pricing. This suggests that the relationship between downstream competition and firm profitability is theoretically ambiguous. In AI supply chains, where downstream firms co-create value through fine-tuning, the dynamics are likely even more complex. Specifically, the strategic interactions between upstream providers and downstream firms make the effects of compute subsidies on their profitability uncertain, as changes in one party's incentives can influence the strategic responses of the other. To address these uncertainties, we pose our second research question:

How do pro-competitive policies and compute subsidies influence the profitability of both foundation model providers and downstream firms with the AI supply chain?

While evaluating the immediate effects of regulatory policies is important, it is equally crucial to recognize how these policies must evolve over time to respond to the shifting dynamics of AI markets. This also represents a key practical challenge for policymakers due to the rapid and continuous evolution of AI technologies (Competition Bureau Canada 2025). A key driver of these changes is the steady decline in compute costs, largely due to advancements in graphics processing units (GPUs), which are essential for fine-tuning AI models (NVIDIA 2024, Diehl 2025). Lower compute costs enable providers to reduce fine-tuning prices, benefiting downstream firms. However, prior research suggests that falling IT costs may intensify competition over product quality, potentially eroding downstream profitability (Demirhan et al. 2005). Consequently, the overall impact of declining compute costs on downstream profitability remains uncertain. In AI supply chains, where downstream firms co-create value through fine-tuning, these dynamics may further influence incentives to invest in quality and, ultimately, consumer surplus. This raises important policy considerations: Regulatory interventions that are effective under current compute cost structures may require adjustment as costs decline in order to maintain their intended effects on consumer surplus and firm incentives. To examine the implications of declining compute costs, we pose our third research question: *How do declining compute costs impact consumer surplus and the profitability of both foundation model providers and downstream firms within the AI supply chain, and how should regulatory policies evolve to ensure welfare-enhancing outcomes?*

1.5. Key Findings

To address these research questions, we develop a game-theoretic model in which a foundation model provider offers both fine-tuning and inference services to two competing downstream firms. These firms, in turn, leverage these services to develop and sell AI-powered products to consumers. Our analysis yields several key insights. First, we show that neither policies promoting price competition in downstream AI markets nor compute subsidies always effectively enhance consumer welfare. Nevertheless, these two policy strategies are complementary, with their effectiveness depending on underlying cost conditions. Specifically, policies promoting price competition in downstream AI markets are effective only when compute or data preprocessing costs are relatively high, whereas compute subsidies are effective only when these costs are relatively low. In contrast, policies promoting quality competition in downstream AI markets always enhance consumer welfare, making it a robust and broadly effective regulatory approach.

Second, contrary to conventional wisdom, we find that policies promoting price competition among downstream firms can increase their profits while reducing the foundation model provider's profits. Moreover, both policies promoting price competition and compute subsidies can, in some cases, generate a "win–win–win" outcome in which all stakeholders benefit. In contrast, policies promoting quality competition consistently

reduce downstream firms' profits while increasing the provider's profits, offering no such "win–win–win" potential.

Third, as GPU technology advances and compute costs decline over time, both consumer surplus and the provider's profits increase, but the downstream firms' profits may decrease. Furthermore, declining compute costs alter the effectiveness of policy interventions. Policies promoting price competition may shift from effective to ineffective, while compute subsidies may transition from ineffective to effective. In contrast, policies promoting quality competition remain consistently effective.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature and situates our work within existing research. Section 3 presents the modeling framework underlying our analysis, while Section 4 develops formal models to examine the optimal decision-making processes of both the foundation model provider and the downstream firms. Section 5 evaluates the effectiveness of various regulatory interventions in AI markets, and Section 6 examines the impact of pro-competitive policies and compute subsidies on the profitability of key stakeholders. Section 7 explores how declining compute costs influence stakeholder profitability, consumer surplus, and the effectiveness of these regulatory policies. Finally, Section 8 concludes with a discussion of theoretical contributions and managerial implications.

2. Literature Review

Our study relates primarily to four streams of literature: (i) AI Regulation, (ii) Software as a Service (SaaS), and (iii) Implications of Declining Information Technology (IT) Costs. Below, we briefly review the relevant literature and highlight our contributions to each stream, with Figure 3 summarizing these works and illustrating the positioning of our study.

2.1. AI Regulation

Our paper contributes to the recently emerging literature on the regulation of AI. Prior work has analyzed the risks posed by AI, emphasizing the urgent need for regulatory oversight. For example, Acemoglu (2021) highlights the broad social and economic harms associated with AI, while Cajueiro and Celestino (2025) offer a comprehensive review of the challenges involved in governing rapidly evolving AI technologies. Empirically, Assad et al. (2024) demonstrate that algorithmic pricing can facilitate collusion and weaken competition, underscoring regulators' concerns about the widespread adoption of AI-powered pricing tools. Another stream of literature proposes or examines specific regulatory approaches. Mahari and Pentland (2024) introduce a Regulation-by-Design paradigm, which integrates regulatory objectives directly into AI system architectures. Other studies investigate the unintended effects of particular interventions. For example, Mohammadi et al. (2024) find that policies designed to improve AI explainability may unintentionally worsen outcomes for both consumers and firms, while Wang et al. (2023) show that policies requiring full transparency of AI algorithms can inadvertently harm end users.

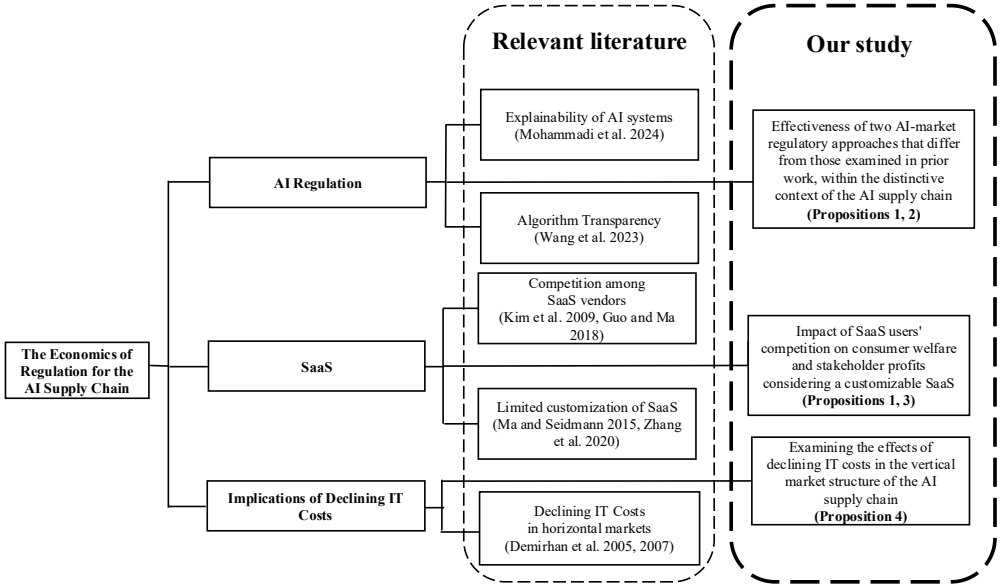
Distinct from these studies, we examine two regulatory approaches that have not yet been analyzed in the AI context, namely pro-competitive policies and compute subsidies. More importantly, we incorporate the multi-layer structure of the AI market, specifically the AI supply chain, a factor that can significantly influence policy effectiveness (Cen et al. 2024b) but has been largely neglected in the literature. By modeling the strategic interactions between upstream foundation model providers and downstream firms, our study provides a more systematic framework for evaluating the impact of regulatory policies within the AI supply chain.

2.2. SaaS

Because fine-tuning and inference services can be viewed as forms of SaaS for AI (Design Gurus 2025), our study is closely related to the literature on SaaS (Yang et al. 2021, Sun et al. 2023, Jiang et al. 2019). Competition among IT resource vendors has been extensively studied in this context. For example, Kim et al. (2009) examine quality competition between SaaS vendors and show that risk-sharing contracts can incentivize vendors to improve product quality. Ma and Kauffman (2014) investigate both price and quality competition, finding that high client switching costs can ultimately drive a vendor out of the market. Similarly, Guo and Ma (2018) and Zhang et al. (2020) analyze the competitive dynamics between perpetual software vendors and SaaS providers, highlighting how factors such as quality improvement rates and network effects shape SaaS competitiveness. While this body of work provides valuable insight into vendor-side competition, it largely overlooks the competitive dynamics among downstream firms that use SaaS products. In particular, little attention has been paid to how competition among downstream users shapes the pricing strategies of SaaS vendors. Our study addresses this gap by examining the interplay between vendor pricing (specifically, by foundation model providers) and downstream market competition. In doing so, we contribute novel insights into how the structure of downstream markets can influence upstream pricing decisions, with implications for both market participants and policymakers.

Prior research has primarily focused on customizable software services (Ma and Seidmann 2015, Zhang et al. 2020) with particular emphasis on modifiable off-the-shelf (MOTS) products. These solutions are typically installed on-premises and offered at fixed prices, rather than through usage-based or on-demand billing, which is a distinctive characteristic of SaaS. In contrast, traditional SaaS models offer limited customization, as multiple customers share the same software instance, thereby restricting access to source code modifications (Sun et al. 2008, Muller et al. 2009, Ma and Kauffman 2014, Guo 2009, Ma and Seidmann 2015, Zhang et al. 2020). While SaaS platforms often include configuration options, these are generally insufficient to accommodate advanced or domain-specific customization needs (Schneier 2015, Zhang et al. 2020, Karanika 2013). The emergence of large-scale AI models introduces a fundamental shift in the nature of service customization. Through fine-tuning, firms can retrain foundation models on proprietary data, enabling parameter adjustments without requiring access to or modification of the underlying source code.

Figure 3 Relevant Literature and the Positioning of Our Study



This capability fosters a new form of co-creation, wherein service quality reflects both the capabilities of the vendor's foundation model and the unique data contributed by downstream firms. This interdependence gives rise to strategic dynamics that diverge significantly from those observed in traditional SaaS settings. Despite these shifts, the strategic implications of customizable SaaS remain underexplored in the existing literature. Our study addresses this gap by examining the evolving interactions between SaaS vendors (i.e., foundation model providers in our context) and downstream firms within this emerging paradigm.

2.3. Implications of Declining IT Costs

Given the continued decline in the costs of AI computing, our paper also contributes to the broader literature on declining IT costs. Prior research in this area has largely focused on the implications of falling IT costs for firm profitability in competitive, consumer-oriented markets. For example, Demirhan et al. (2005, 2007) examine strategic IT investments and show that whether firms can benefit from declining IT costs depends on factors such as price sensitivity and user switching costs. Similarly, Bohlmann et al. (2002) analyze the sequential entry of firms into downstream markets and empirically show that the negative effect of declining IT costs on the profit of the pioneer is more pronounced in product categories where quality is a key differentiator.

In contrast to prior work, our study examines the effects of declining IT costs within a vertical market structure, specifically the AI supply chain that includes upstream providers, downstream firms, and end consumers. Whereas earlier research has primarily focused on horizontal market settings involving interactions only between downstream firms and consumers, our model captures the strategic implications of vertical interactions across the supply chain. Moreover, while prior studies typically attribute IT costs to downstream firms, our analysis highlights the strategic impact of declining compute costs incurred by upstream providers.

3. Modeling Framework

We analyze a market scenario involving a foundation model provider (hereafter referred to as the *provider*) and two competing downstream firms (denoted by *firm i*, where $i \in \{1, 2\}$). To ground the analysis in a real-world example, consider law firms such as Harvey and Thomson Reuters, which are prominent competitors in the provision of AI-powered legal assistants (Fisher 2024). Both firms have fine-tuned OpenAI's GPT model to develop the legal assistants. This fine-tuning process enables the legal assistants to deliver highly relevant and accurate results, thereby enhancing service quality and contributing significantly to firm profitability (OpenAI 2024a). Such scenarios, where competing firms rely on the same foundation model provider, are increasingly common in practice (Shaver 2025).

Next, we define the profit functions of each stakeholder and outline the sequence of the game.

3.1. Foundation model provider

To support firms in leveraging the foundation model, the provider offers two key services: fine-tuning and inference. Both services are essential for firms to deliver their AI-powered products, such as AI-driven legal assistants, to end users, who may be individuals or enterprises.

In the fine-tuning process, firms supply domain-specific data to the provider, who utilizes its computing infrastructure to adapt the foundation model, thereby enhancing its domain-specific performance. According to OpenAI (2023), the performance of a fine-tuned model improves with the volume of data provided. To quantify this relationship, we define a *data unit* as the amount of data required to improve model quality by one unit (e.g., 1,000 tokens). Additionally, we define *data volume* as the volume of data usage for fine-tuning, denoted as V_i. Hence, if firm i provides V_i data units for fine-tuning, the corresponding improvement in model quality is V_i. Therefore, the quality of the fine-tuned model can be represented as $q_i = \alpha + V_i$, where α denotes the base quality of the foundation model. Since fine-tuning is performed on the provider's computing infrastructure, it incurs rising computational costs as the volume of training data increases. These costs are modeled using the quadratic expression $c_F V_i^2$, where c_F is the cost coefficient associated with fine-tuning. The convex cost structure arises because the fine-tuning of large-scale AI models depends on distributed computing architectures (Yanbc 2024, Insights 2024). As the volume of data increases, additional GPU nodes are required, leading to greater communication overhead between GPUs. This, in turn, creates computational bottlenecks that reduce GPU efficiency and cause costs to grow non-linearly (Narayanan et al. 2021, Zhao et al. 2021). To provide fine-tuning services, the provider sets a per-unit price of p_F for the data used in this process. For example, OpenAI charges \$6 per million tokens[2] for fine-tuning on its Davinci model, part of the GPT series (OpenAI 2024b). Consequently, firm i pays a total amount of $p_F V_i$ to the provider.

The fine-tuned model is deployed and hosted on the provider's computing infrastructure. To access the model, firms utilize the provider's inference service. During inference, firms submit queries to the fine-tuned

model, which processes the input, generates responses, and returns the results. The provider typically charges firms for inference on a pay-per-use basis, with a price of p_I per unit of data usage, which includes both the input and output generated by the model (e.g., OpenAI charges $12 per million tokens for inference on its Davinci model) (OpenAI 2024b). Since downstream firms develop AI-powered products based on the fine-tuned model, each consumer interaction with the product triggers use of the inference service. For example, each query submitted by a lawyer to an AI-powered legal assistant invokes the fine-tuned model to generate a response, thereby incurring inference usage. For simplicity, we assume that answering a consumer's query requires one unit of data usage for inference. Therefore, the total data usage for inference equals the total number of consumers, denoted by d_i. Consequently, firm i pays $p_I d_i$ to the provider for inference services.

The objective of the foundation model provider is to maximize total profits by strategically setting prices for both fine-tuning and inference services. The provider's profit function is given by:

$$\pi_M = \sum_{i=1}^{2} \left(p_I d_i + p_F V_i - c_F V_i^2 \right). \tag{1}$$

In this expression, the first term ($p_I d_i$) represents revenue generated from the inference service, where d_i denotes the inference demand for firm i. The second term ($p_F V_i$) corresponds to revenue from fine-tuning services, while the third term ($c_F V_i^2$) captures the computational costs associated with fine-tuning.

3.2. Firms

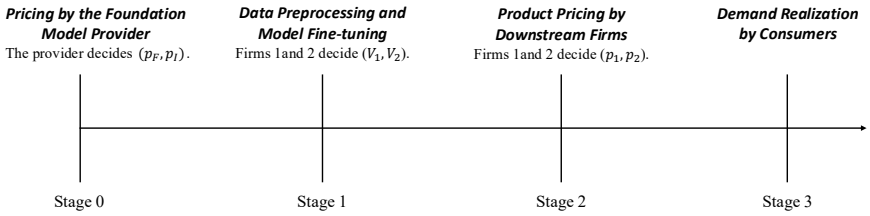
Firms must preprocess their domain-specific data to address potential issues in the raw data and ensure its suitability for fine-tuning (Huizenga 2025). For example, de-duplication, one of the key preprocessing step, involves identifying and removing duplicate data points (Huizenga 2025). Although essential, data preprocessing represents a significant burden for firms. For instance, Thomson Reuters reportedly invested 4,000 hours of effort from a multidisciplinary team of lawyers, domain experts, and AI engineers to curate and refine 30,000 legal questions for fine-tuning (Kardach 2024). The workload associated with preprocessing increases with the volume of data used for fine-tuning (V_i), as larger preprocessed datasets require more raw data and greater processing complexity. For example, the de-duplication process involves pairwise comparisons and thus exhibits quadratic computational complexity with respect to the dataset size (Munappy et al. 2022). Accordingly, we model the cost of data preprocessing as $c_V V_i^2$.

To model price and quality competition in the downstream market, we build on prior research (Wang et al. 2022, Tsay and Agrawal 2000) by assuming that aggregate consumer demand is linearly influenced by both price and product quality. We further assume that the quality of an AI-powered product corresponds to the quality of the fine-tuned model, denoted by q_i. Specifically, consumer demand for firm i is given by:

$$d_i = 1 - p_i - \theta_p(p_i - p_j) + q_i + \theta_q(q_i - q_j),$$

where $i, j \in \{1, 2\}$ and $i \neq j$. Under this formulation, lowering a firm's price or improving its product quality increases its own demand while reducing that of its competitor, *ceteris paribus*. The parameters $\theta_p > 0$ and

Figure 4 Model Timeline



$\theta_q > 0$ capture the intensity of price and quality competition, respectively, in the consumer market. As firms generate revenue ($p_i d_i$) by selling AI-powered products to consumers, they simultaneously incur inference costs ($p_I d_i$) paid to the provider when consumers interact with the underlying model.

Accordingly, firm i's profit equals its sales revenue minus inference costs, fine-tuning costs, and data preprocessing costs. The resulting profit function for firm i is given by:

$$\pi_i = (p_i - p_I)d_i - p_F V_i - c_V V_i^2. \tag{2}$$

3.3. Model Timeline

The game unfolds over four stages, as illustrated in Figure 4, which depicts the interactions among key stakeholders, namely the foundation model provider, the downstream firms, and the consumers. In Stage 0, the provider sets the fine-tuning price (p_F) and inference price (p_I), both of which are observable to the firms. In Stage 1, Firms 1 and 2 simultaneously choose their respective data volumes for fine-tuning (V_i, where $i \in 1, 2$), preprocess their domain-specific data, and fine-tune the foundation model accordingly. Once fine-tuning is complete, the firms develop AI-powered products based on the fine-tuned model, resulting in product quality denoted by q_i. In Stage 2, after observing each other's product quality, the firms simultaneously set their product prices (p_i) and offer their AI-powered products to consumers. Finally, in Stage 3, consumer demand is realized. This sequential structure, in which quality decisions precede pricing, aligns with established models of quality and price competition in the literature (Banker et al. 1998, Demirhan et al. 2005). The underlying rationale is that quality decisions are typically more difficult to modify than pricing decisions; thus, firms often determine product quality prior to setting prices (Banker et al. 1998, Demirhan et al. 2005).

4. Equilibrium Analysis

In this section, we employ backward induction to derive the subgame-perfect Nash equilibrium (SPNE) for the players' actions at each stage of the game, highlighting our key findings. Following the sequence outlined in Figure 4, we begin by determining the product prices and data volumes chosen by the downstream firms, given the fine-tuning and inference pricing decisions made by the foundation model provider. We then proceed to analyze the provider's equilibrium choices with respect to these upstream decisions.

4.1. Downstream Firms' Decisions

In this subsection, we focus on the subgame starting at Stage 1, where the provider's inference price (p_I) and fine-tuning price (p_F) are treated as given. Using backward induction, we begin our analysis at Stage 2, where downstream firms determine their product prices to maximize their respective profit functions, as specified in Equation (2). In this stage, firms take as given the realized product qualities, along with the previously set prices for fine-tuning and inference. We then proceed to derive firm i's best-response function for Stage 2 as:

$$p_i(V_i, V_j, p_I) = \frac{p_I(1+\theta_p)(2+3\theta_p) + 2(1+\alpha+V_i+\theta_q(V_i-V_j)) + \theta_p(3+3\alpha+2V_i+V_j+\theta_q(V_i-V_j))}{(2+\theta_p)(2+3\theta_p)}, \tag{3}$$

where $i \neq j \in \{1,2\}$.

By substituting the above expressions for prices into the firms' profit functions from Equation (2), we derive the firms' profits in Stage 1. At this stage, each firm selects its data volume (V_i) for fine-tuning to maximize its profit. Solving the first-order condition yields the SPNE data volume for firm i (denoted as $V_i(p_F, p_I)$).[3] The corresponding product prices ($p_i(p_F, p_I)$) can be obtained by substituting $V_i(p_F, p_I)$ into Equation (3). The following lemma summarizes the firm's equilibrium decisions in this subgame at Stage 1.

LEMMA 1 (**Equilibrium Decisions of the Firms in the Subgame**). *Given the inference price p_I and fine-tuning price p_F set by the foundation model provider, the equilibrium data volumes and product prices for Firm i ($i \in \{1,2\}$) in the subgame at Stage 1 are as follows:*

$$V_i(p_I, p_F) = \frac{2(1+\theta_p)(1+\alpha-p_I)(2(1+\theta_q)+\theta_p(2+\theta_q)) - (2+\theta_p)^2(2+3g)p_F}{2c_V(2+\theta_p)^2(2+3\theta_p) - 2(1+\theta_p)(2(1+\theta_q)+\theta_p(2+\theta_q))}, \tag{4}$$

$$p_i(p_I, p_F) = \frac{\begin{array}{c}2c_V(2+\theta_p)(2+3\theta_p)(1+\alpha) + 2p_I(1+\theta_p)(c_V(2+\theta_p)(2+3\theta_p) \\ -\theta_p\theta_q - 2(1+\theta_p+\theta_q)) - (2+\theta_p)(2+3\theta_p)p_F\end{array}}{2c_V(2+\theta_p)^2(2+3\theta_p) - 2(1+\theta_p)(2(1+\theta_q)+\theta_p(2+\theta_q))}. \tag{5}$$

Accordingly, we can rewrite Firm i's profit function by substituting the above prices and data volumes into Equation (2), yielding the following expression:

$$\pi_i = \big(p_i(p_I, p_F) - p_I\big)\big(1+\alpha+V_i(p_I, p_F) - p_i(p_I, p_F)\big) - p_F V_i(p_I, p_F) - c_V V_i^2(p_I, p_F) \tag{6}$$

As established in Lemma 1, the equilibrium data volumes and product prices in the subgame are influenced by the intensity of downstream competition, captured by the parameters θ_p (price competition) and θ_q (quality competition). Specifically, an increase in price competition leads to reductions in both data volumes and product prices (i.e., $\frac{\partial V_i(p_I,p_F)}{\partial \theta_p} < 0$, $\frac{\partial p_i(p_I,p_F)}{\partial \theta_p} < 0$). In contrast, an increase in quality competition results in higher data volumes and product prices (i.e., $\frac{\partial V_i(p_I,p_F)}{\partial \theta_q} > 0$, $\frac{\partial p_i(p_I,p_F)}{\partial \theta_q} > 0$). The underlying mechanism is as follows: As price competition intensifies, firms are compelled to lower their product prices (i.e., $\frac{\partial p_i(p_I,p_F)}{\partial \theta_p} < 0$), thereby diminishing the marginal returns to quality improvements through increased

consumer demand. Consequently, firms decrease the volume of data used for fine-tuning (i.e., $\frac{\partial V_i(p_I,p_F)}{\partial \theta_p} < 0$). Conversely, heightened quality competition incentivizes firms to enhance product quality by increasing data volumes for fine-tuning (i.e., $\frac{\partial V_i(p_I,p_F)}{\partial \theta_q} > 0$). This enhancement in quality enables firms to command higher prices (i.e., $\frac{\partial p_i(p_I,p_F)}{\partial \theta_q} > 0$), as consumers are willing to pay a premium for superior product quality resulting from more extensive data use. We now proceed to examine the provider's decision-making in Stage 0.

4.2. Foundation Model Provider's Decisions

In this subsection, we analyze the foundation model provider's equilibrium pricing decisions. To this end, we substitute the SPNE prices and data volumes into the provider's profit function, as specified in Equation (1). This allows us to rewrite the provider's profit function as:

$$\pi_M = \sum_{i=1}^{2} p_I d_i(p_I, p_F) + p_F V_i(p_I, p_F) - c_F V_i^2(p_I, p_F), \tag{7}$$

where $d_i(p_I, p_F) = 1 + \alpha + V_i(p_I, p_F) - p_i(p_I, p_F)$. The provider chooses the inference price (p_I) and the fine-tuning price (p_F) to maximize profit, such that $\pi_M^* = \max_{p_I, p_F} \pi_M$.

By solving the corresponding first-order conditions, we obtain the equilibrium inference and fine-tuning prices, denoted by p_I^* and p_F^*, respectively.[4] These results are summarized in the following lemma.

LEMMA 2 (**Equilibrium Pricing Decisions of the Provider**). *In equilibrium, the foundation model provider sets the inference and fine-tuning prices as follows:*

$$p_I^* = \frac{\begin{array}{c}2(1+\alpha)(c_F(2+\theta_p)^3(2+3\theta_p)^2 + 2c_V(2+\theta_p)^3(2+3\theta_p)^2 \\ -(1+\theta_p)(2+2\theta_p+2\theta_q+\theta_p\theta_q)(8+12\theta_p+3\theta_p^2+2(2+\theta_p)\theta_q))\end{array}}{4c_F(2+\theta_p)^3(2+3\theta_p)^2 + 8c_V(2+\theta_p)^3(2+3\theta_p)^2 - (1+\theta_p)(8+12\theta_p+3\theta_p^2+2(2+\theta_p)\theta_q)^2},$$

$$p_F^* = \frac{2(1+\theta_p)(2+\theta_p)(2+3\theta_p)(1+\alpha)(4c_F+4c_F\theta_p-4c_V\theta_p-3c_V\theta_p^2+2(c_F+c_V)(2+\theta_p)\theta_q)}{4c_F(2+\theta_p)^3(2+3\theta_p)^2 + 8c_V(2+\theta_p)^3(2+3\theta_p)^2 - (1+\theta_p)(8+12\theta_p+3\theta_p^2+2(2+\theta_p)\theta_q)^2}.$$

Accordingly, the equilibrium data volumes and product prices, denoted by V_i^* and p_i^*, are obtained by substituting p_I^* and p_F^* into Equations (4) and (5), respectively. As shown in Lemma 2, the provider's equilibrium prices depend on the downstream market parameters (θ_p, θ_q). We therefore examine how the intensity of downstream price and quality competition affects the provider's optimal pricing decisions, as summarized in the following lemmas.

LEMMA 3 (**Effects of Downstream Competition on the Fine-Tuning Price**).

a. The fine-tuning price deceases with the intensity of downstream price competition (i.e., $\frac{dp_F^*}{d\theta_p} < 0$).

b. The fine-tuning price increases with the intensity of downstream quality competition (i.e., $\frac{dp_F^*}{d\theta_q} > 0$).

Lemma 3 demonstrates that the fine-tuning price decreases with downstream price competition (θ_p) and increases with downstream quality competition (θ_q). As price competition intensifies, the firms reduce the

data volumes used for fine-tuning, as explained following Lemma 1. This reduction not only lowers fine-tuning revenue but also diminishes inference revenue due to the resulting degradation in product quality. Consequently, the provider is incentivized to decrease the fine-tuning price to stimulate data volume, which leads to $\frac{dp_F^*}{d\theta_p} < 0$. In contrast, as quality competition intensifies, the firms increase data volume, as discussed after Lemma 1, prompting the provider to raise the fine-tuning price.

The next lemma examines how downstream price and quality competition affects the inference price.

LEMMA 4 **(Effects of Downstream Competition on Inference Price)**.

a. The inference price increases with the intensity of downstream price competition ($\frac{dp_I^*}{d\theta_p} > 0$) when $c_F + 2c_V > c_1$ and decreases otherwise.

b. The inference price decreases with the intensity of downstream quality competition ($\frac{dp_I^*}{d\theta_q} < 0$) when $c_F + 2c_V > c_2$ and increases otherwise.

For brevity, the expressions for c_1 and c_2 are provided in the Online Appendix.

In contrast to the fine-tuning price, both downstream price and quality competition have non-monotonic effects on the inference price, as indicated by Lemma 4. The outcome arises because price competition influences inference prices in two distinct ways. First, intensified price competition reduces product prices, which, in turn, increases inference demand, prompting the provider to raise the inference price. Second, lower product prices, resulting from heightened price competition, reduce the marginal benefits of product quality, leading the firms to invest less in quality improvement. This is achieved by decreasing the data volume used for fine-tuning. The reduction in data volume diminishes the provider's revenue, prompting a decrease in the inference price, which increases the marginal benefit of quality improvement and thereby stimulates data volume. However, the magnitude of this effect depends on the firms' marginal cost of data volume. Notably, it can be shown that when c_F or c_V is relatively high, specifically when $c_F > c_1 - 2c_V$ or $c_V > (c_1 - c_F)/2$ (equivalent to the conditions specified in Lemma 4.a), the marginal cost of data volume is also high. In such cases, stimulating a higher data volume through a reduction in inference price becomes more difficult. Consequently, the provider may not lower the inference price significantly, leading to an overall increase in the inference price. Conversely, when the marginal cost of data volume is relatively low, the second effect dominates. A similar rationale explains the effects of increased quality competition on the inference price.

5. Consumer Surplus and Policy Implications

With the expanding use of AI-enabled products, consumers welfare face significant risks, such as the difficulty of assessing the quality of AI-enabled products (Ngo 2025). In light of these concerns, policymakers are increasingly focused on preserving consumer welfare in the context of AI(Robinson 2025). However, the complex interdependencies between upstream providers and downstream markets, detailed in Section

4, underscore the regulatory challenges inherent to the AI supply chain (Cen et al. 2024b). This section addresses these issues by evaluating the effectiveness of existing regulatory frameworks and offering insights for selecting effective policy instruments. We focus on two prominent approaches: promoting downstream market competition and providing compute subsidies (Competition Bureau 2024, Wu 2024). Our analysis proceeds in two parts: first, we assess the impact of pro-competitive policies on consumer surplus; second, we evaluate the effectiveness of compute subsidy policies. We conclude by discussing the conditions under which each policy instrument most effectively enhances consumer surplus.

5.1. Pro-Competitive Policies in Downstream Markets

For policymakers, pro-competitive regulation has long been a key strategy to enhance consumer welfare, primarily by lowering prices and improving product quality (Tucker 2024). To lower prices, a major component of pro-competitive policies aims to promoting price competition in the consumer market (hereafter referred to as the *pro-price-competitive policy*), which can be achieved by approaches such as enhancing price transparency or facilitating price comparisons across products. One prominent instance is the Federal Trade Commission's (FTC) Rule on Unfair or Deceptive Fees. This rule mandates that businesses disclose the total price upfront, prohibiting hidden or misleading fees.(FTC 2025d). In a recent case, Ticketmaster, a leading online ticketing platform, was accused by the FTC of violating this rule by failing to provide full price transparency. As a result, the company was required to adjust its pricing system to ensure that consumers could see the total cost before completing a purchase (Boggs 2025). Similarly, the U.S. Federal Communications Commission (FCC) has implemented regulations requiring Internet service providers to display their base monthly prices on standardized labels. This rule aims to improve consumer decision-making by making it easier for individuals to compare prices across providers, thus enhancing market transparency and fostering more competitive pricing (FCC 2025).

Another key component of pro-competitive policies focuses on fostering quality competition in consumer markets (hereafter referred to as the *pro-quality-competitive policy*), often through increased transparency around product quality. A notable example of this is the FTC's *Trade Regulation Rule on the Use of Consumer Reviews and Testimonials*, which aims to prevent deceptive practices like fake reviews and the suppression of negative feedback (FTC 2025a). In one case, the FTC alleged that Fashion Nova misrepresented its product reputation by concealing customer reviews with ratings below four stars, thereby misleading consumers about the true quality of its products (FTC 2025c). Similarly, the European Union's Consumer Rights Act mandates that businesses disclose the total price of digital content or services, including AI-based products, and prohibits hidden fees. Together, these policies promote greater transparency around both price and quality, which in turn fosters more intense competition in the downstream segment of the AI supply chain. As economic literature suggests, increased market transparency tends to heighten competition, ultimately benefiting consumers.

In the context of the AI supply chain, policymakers are increasingly advocating for the development of similar pro-competitive policies (Competition Bureau 2024). The FTC, for example, has issued guidance that prohibits companies from exaggerating or misrepresenting the capabilities of their AI-based products (Leblanc et al. 2025). Companies found guilty of overstating their AI products, such as DoNotPay, have faced FTC complaints and proposed settlements for deceptive claims (FTC 2025b). Similarly, the European Union's *Consumer Rights Act* mandates that businesses disclose the total price of digital content or services, including AI-based products, and prohibits hidden fees (CCPC 2025). Together, these policies promote greater transparency around both price and quality, which in turn foster more intense competition in the downstream segment of the AI supply chain. [1]

Recent research and policy analyses highlight that heightened competition in product markets can benefit consumers by driving down prices or encouraging quality improvements (Government of Canada 2024, Tyagi 1999, Wu and Zhu 2022). However, existing literature also point to potential downsides, suggesting that intensified price competition may diminish firms' incentives to invest in product quality (Tsay and Agrawal 2000), ultimately leading to lower product quality and reduced consumer surplus. A similar concern arises in the context of quality competition, where research indicates that aggressive competitive pressures may lead to a decline in product quality(Armstrong and Chen 2009). In the context of the AI supply chain, it remains unclear which of these opposing effects dominates, raising questions about whether pro-competitive policies in downstream markets ultimately enhance or reduce consumer surplus.

To address this issue, we first define consumer surplus, denoted as CS, within our model framework. Following the literature (Singh and Vives 1984, Wang et al. 2024), we consider a representative consumer, and derive CS from the corresponding utility function (i.e., $U(d_i, d_j) = \sum_{i=1}^{2} \mu_i d_i - \frac{\beta_i d_i^2 + \gamma d_i d_j}{2} - p_i d_i$, where the first term calculates the direct benefit of consuming product, the second term calculates the diminishing return of consuming products, and the last term calculates the payment made by consumers). The resulting expression of CS and its derivation are provided in the Online Appendix. In equilibrium, this expression simplifies to:

$$CS = \sum_{i=1}^{2} \frac{1}{2}(1 + \alpha + V_i^* - p_i^*)^2, \tag{8}$$

which indicates that consumer surplus is jointly determined by the product qualities ($q_i^* = \alpha + V_i^*$) and product prices p_i^*.

Based on Equation (8), we analyze the effects of pro-price-competitive and pro-quality-competitive policies in the downstream market on consumer surplus. The results are summarized in in the following proposition.

PROPOSITION 1 **(Impact of Pro-competitive Policies on Consumer Surplus)**.

[1] Studies show that greater market transparency often leads to increased competition, ultimately benefiting consumers (Schultz 2009).

a. As price competition in downstream markets intensifies, consumer surplus decreases if $c_F + 2c_V < \tilde{c}$, and increases otherwise.

b. As quality competition in downstream markets intensifies, consumer surplus increases.

The explicit expression for $\tilde{c}$ is algebraically complex and is thus provided in the Online Appendix.

As shown in Proposition 1a, while pro-competitive policies may intensify price competition and reduce product prices, consumers surplus can paradoxically decline when compute costs (i.e., c_F) or data preprocessing costs (i.e., c_V) are relatively low. This counterintuitive result can be explained as follows.

As price competition intensifies, firms are incentivized to lower their product prices, which benefits consumers (i.e., $\frac{dp_i^*}{d\theta_p} < 0$). However, the impact on product quality is more nuanced, as it can move in either direction due to the co-creation dynamics inherent in the AI supply chain. The direct effect of increased price competition is a reduction in product quality (i.e., $\frac{\partial V_i(p_I,p_F)}{\partial \theta_p}\big|_{p_I^*,p_F^*} < 0$): Lower prices reduce the marginal benefit of offering higher quality, thus diminishing firms' incentives to invest in product quality.

In contrast, the indirect effect operates through the provider's pricing response (i.e., $\big(\frac{\partial V_i(p_I,p_F)}{\partial p_I}\frac{dp_I^*}{d\theta_p} + \frac{\partial V_i(p_I,p_F)}{\partial p_F}\frac{dp_F^*}{d\theta_p}\big)|_{p_I^*,p_F^*}$). Specifically, as price competition intensifies, the provider lowers the fine-tuning price (i.e., $\frac{dp_F^*}{d\theta_p} < 0$), while the inference price may either rise or fall (i.e., $\frac{dp_I^*}{d\theta_p}$ can take either sign), as shown in Lemmas 3 and 4. A higher inference price reduces the marginal benefit of product quality, leading to a decrease in data volumes ($\frac{\partial V_i(p_I,p_F)}{\partial p_I}\big|_{p_I^*,p_F^*} < 0$), while a lower fine-tuning price reduces the marginal cost of data volume, thus increasing data volumes ($\frac{\partial V_i(p_I,p_F)}{\partial p_F}\big|_{p_I^*,p_F^*} < 0$). Even if the inference price rises and suppresses data volumes, the effect of the reduction in the fine-tuning price dominates, resulting in an overall increase in product quality ($\big(\frac{\partial V_i(p_I,p_F)}{\partial p_I}\frac{dp_I^*}{d\theta_p} + \frac{\partial V_i(p_I,p_F)}{\partial p_F}\frac{dp_F^*}{d\theta_p}\big)\big|_{p_I^*,p_F^*} > 0$). Taken together, intensified price competition indirectly encourages higher product quality by enhancing the provider's co-creation incentives through lower the fine-tuning price (and sometimes the inference price). However, as explained below, when the coefficients of data preprocessing costs and compute costs (i.e., c_V and c_F) are relatively low, this indirect effect is dominated by the direct effect. In this case, the net result is a decline in product quality, which in turn negatively impacts consumer surplus.

When c_V increases, the firms' marginal cost of quality creation rises, which reduces the direct effect on quality creation (i.e., $\frac{d}{dc_V}\big(\frac{\partial V_i(p_I,p_F)}{\partial \theta_p}\big|_{p_I^*,p_F^*}\big) > 0$). A similar effect arises when c_F increases: Elevated compute costs lead to a higher fine-tuning price, which in turn increases the marginal cost of quality improvement, further diminishing the direct effect (i.e., $\frac{d}{dc_F}\big(\frac{\partial V_i(p_I,p_F)}{\partial \theta_p}\big|_{p_I^*,p_F^*}\big) > 0$). Thus, when either c_V or c_F is relatively high, the negative direct effect of intensified price competition on quality creation is dominated, leading to an increase in product quality and, consequently, consumer surplus as θ_p increases. In contrast, when both c_V and c_F are relatively low (i.e., $c_F + 2c_V < \tilde{c}$), the direct effect on product quality dominates, leading to a decline in product quality and, consequently, consumer surplus as θ_p increases.

Proposition 1b further demonstrates that consumers always benefit as quality competition in the downstream market intensifies. At first glance, this may seem intuitive, as heightened quality competition encourages firms to enhance product quality (i.e., $\frac{dV_i^*}{d\theta_q} > 0$). However, our analysis uncovers an additional, more nuanced mechanism: In certain cases, intensified quality competition also leads to lower product prices (i.e., $\frac{dp_i^*}{d\theta_q} < 0$ if $c_F + 2c_V > \frac{(1+\theta_p)(3\theta_p^2+4(2+\theta_q)+2\theta_p(6+\theta_q))^2}{4(2+\theta_p)(2+3\theta_p)(\theta_p^2+2(1+\theta_p)(2+\theta_p)\theta_q)}$), which in turn benefits consumers. This outcome results from indirect effects driven by the strategic pricing behavior of the foundation model provider within the AI supply chain. Specifically, as quality competition intensifies, the provider may lower the inference price (as established in Lemma 4), which in turn incentivizes firms to reduce their product prices. Thus, the increase in consumer surplus from intensified quality competition is not solely driven by higher product quality; it can also stem from lower product prices. This challenges the conventional view that higher-quality products always command higher prices, as suggested in the literature (e.g., Tirole 1988).

5.2. Compute Subsidies

In addition to pro-competitive policies, subsidizing access to computing power (hereafter referred to as *compute subsidies*) is a widely used regulatory tool employed by policymakers. By reducing compute costs, these subsidies incentivize firms to leverage more data for model fine-tuning, ultimately improving product quality and increasing consumer surplus (Sastry et al. 2024). In China, the government has implemented this approach within the AI supply chain. For example, the Beijing municipal government allocates annual compute subsidies totaling 100 million RMB to firms involved in large-scale AI model services. These subsidies can offset up to 30% of computing expenses, including those associated with fine-tuning (Wu 2024). Similarly, the Shenzhen municipal government has introduced an annual compute subsidy program worth 500 million RMB, covering up to 50% of compute costs. Comparable initiatives have also been launched by the Shanghai and Chengdu municipal governments (Wu 2024).

To analyze the impact of this policy on consumer surplus, we consider the scenario where policymakers offer compute subsidies that reduce the used compute costs by a proportion x. We define x as the subsidy rate, representing the share of compute cost covered by policymakers. Under this policy, the foundation model provider incurs fine-tuning costs of $c_F(1-x)V_1^2 + c_F(1-x)V_2^2$, rather than the full cost $c_F V_1^2 + c_F V_2^2$. Let V_1' and V_2' denote the resulting equilibrium data volumes, and CS' represent the corresponding consumer surplus under the compute subsidy. The total subsidy expenditure under this policy is given by $xc_F V_1'^2 + xc_F V_2'^2$. We then compare the consumer surplus with and without the policy, and present the results in the following lemma.

LEMMA 5 (**Impact of Compute Subsidies on Consumer Surplus**). *When compute subsidies are provided, consumer surplus increases compared to the scenario without them (i.e.,* $\forall x \in [0,1]$, $CS'(x) > CS(x)$*).*

Lemma 5 shows that, although compute subsidies are granted to the foundation model provider, they ultimately benefit consumers. This can be explained as follows: When policymakers subsidize compute

costs, the foundation model provider faces a lower marginal compute cost, which leads to a reduction in the fine-tuning price. The lower fine-tuning price encourages firms to increase their data volumes for fine-tuning, thereby improving product quality and enhancing consumer surplus.

Although compute subsidies increase consumer surplus, it is essential to evaluate whether the policy is cost-effective, especially given the substantial public funds involved. To do so, we adopt the framework from the literature (Anderson et al. 2022) and define *cost-effectiveness* (CE) as the net benefits of compute subsidies, which are measured by the difference between the incremental consumer surplus gained from the compute subsidies and the subsidy expenditure, and can be expressed as $CS' - CS - xc_F V_1'^2 - xc_F V_2'^2$. A positive value indicates that the policy is cost-effective in enhancing consumer surplus. We formally analyze this effectiveness metric and present the results in the following proposition.

PROPOSITION 2 **(Cost-Effectiveness of Compute Subsidies)**. *Compute subsidies are cost-effective only when $c_F + 2c_V < \hat{c}$. The explicit expression for $\hat{c}$ is algebraically complex and is thus provided in the Online Appendix.*

Although consumer surplus increases with compute subsidies (as shown in Lemma 5), Proposition 2 indicates that such subsidies are cost-ineffective in enhancing consumer surplus when compute and data preprocessing costs are relatively high. This counter-intuitive result can be explained as follows: While compute subsidies improve product quality and thus enhance consumer surplus (as discussed following Lemma 5), they also incur substantial costs. When these costs outweigh the increase in consumer surplus, the subsidies become cost-ineffective. When the coefficient of data preprocessing costs (c_V) is relatively high, firms face substantial marginal costs in expanding data volumes for fine-tuning, limiting their ability to improve product quality even with subsidized compute costs. As a result, the increase in consumer surplus is smaller than the subsidy expenditure. Similarly, when the coefficient of compute costs (c_F) is high, the provider still faces significant marginal costs despite the subsidy and is thus incentivized to maintain relatively high fine-tuning prices. This results in high marginal costs for firms attempting to improve product quality, thereby limiting the policy's cost-effectiveness. Therefore, when c_F or c_V is relatively high, compute subsidies are cost-ineffective in improving consumer surplus.

As shown in Proposition 2, compute subsidies can be cost-effective when compute costs and data preprocessing costs are relatively low. Specifically, when $c_F + 2c_V < \hat{c}$, there exists a subsidy rate (i.e., x) such that $CE > 0$. Thus, whether these subsidies are cost-effective or not also depends on the choice of subsidy rate (x). Therefore, we examine the relationship between cost-effectiveness and the subsidy rate for computing power under the condition ($c_F + 2c_V < \hat{c}$), and summarize our findings in the following corollary.

As shown in Proposition 2, compute subsidies can be cost-effective when compute costs and data preprocessing costs are relatively low. Specifically, when $c_F + 2c_V < \hat{c}$, there exists a subsidy rate (i.e., x) such that $CE > 0$. This implies that whether these subsidies are cost-effective also depends on the choice

of the subsidy rate. Thus, we examine the relationship between cost-effectiveness and the subsidy rate for computing power under the condition ($c_F + 2c_V < \hat{c}$), and summarize our findings in the following corollary.

COROLLARY 1 **(Relationship between Cost-Effectiveness and Subsidy Rate for Compute Costs)**. *As the subsidy rate for compute costs (i.e., x) increases, cost-effectiveness improves if $x < x'$, and decreases if $x > x'$. Additionally, if $x > x''$ (with $x'' > x'$), compute subsidies become cost-ineffective; otherwise, they become cost-effective.*

For brevity, the explicit expressions for x' and x'' are provided in the Online Appendix.

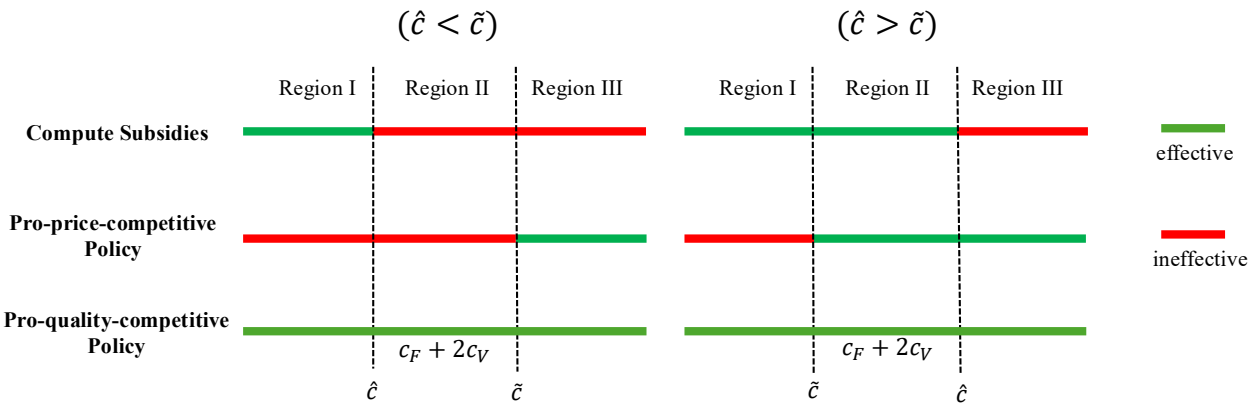
Corollary 1 demonstrates that the net benefits of compute subsidies do not increase monotonically with the subsidy rate x. Instead, the net gain in consumer surplus relative to the subsidy costs initially rises and then falls as x increases, with cost-effectiveness peaking at x'. This non-monotonic pattern can be explained by the fine-tuning behavior of downstream firms: At low subsidy rates, increasing x encourages firms to expand data volumes for fine-tuning, which improves product quality, increases consumer surplus, and enhances the cost-effectiveness. This explains the initial upward trend for values of x below x'. However, once the subsidy rate exceeds this threshold (i.e., $x > x'$), the provider responds by lowering the fine-tuning price excessively, which incentivizes firms to engage in more aggressive fine-tuning. Although this improves product quality and further increases consumer surplus, it also results in a substantial increase in data volume, thereby driving up subsidy costs ($xc_F V_1'^2 + xc_F V_2'^2$). As a result, the gains in consumer surplus are eventually outweighed by the increase in subsidies when $x > x'$, causing the cost-effectiveness of compute subsidies to decline. If the subsidy rate x increases beyond x'', cost-effectiveness may even turn negative. Therefore, under the condition $c_F + 2c_V < \hat{c}$, compute subsidies remain cost-effective only if the subsidy rate is not too high (i.e., $x < x''$).

Proposition 2 offers valuable regulatory insights for policymakers. The findings suggest that compute subsidies must be implemented with caution, as their effectiveness in enhancing consumer surplus depends heavily on the cost environment. Specifically, when data preprocessing costs or compute costs are relatively high, such subsidies may yield limited or even negative returns. Additionally, to ensure the effectiveness of compute subsidies, it is also crucial to carefully calibrate the subsidy rate. If the rate is set too high, it could encourage downstream firms to overinvest in fine-tuning, leading to disproportionately high subsidy expenditures and diminishing, or even negative, returns in terms of net consumer surplus gains.

5.3. Comparing Policies

Building on the previous analysis, this subsection contrasts each policy's effectiveness in enhancing consumer surplus across different cost regions. Unlike compute subsidies, which impose a direct cost on the policymaker, the pro-competitive policy does not involve any government expenditure. Thus, the effectiveness of the pro-competitive policy is assessed based solely on its ability to increase consumer surplus. In contrast, compute subsidies are deemed effective only when the increase in consumer surplus exceeds the costs incurred in

Figure 5 Policy Effectiveness under Different Cost Regions



providing the subsidy Figure 5 illustrates the effective and ineffective policy regimes for compute subsidies, the pro-price-competitive policy, and the pro-quality-competitive policy. We partition the cost space into three distinct regions. When compute and data preprocessing costs are relatively low ($c_F + 2c_V < \min\{\tilde{c}, \hat{c}\}$), the cost condition falls into Region I. When these costs are moderate ($\min\{\tilde{c}, \hat{c}\} < c_F + 2c_V < \max\{\tilde{c}, \hat{c}\}$), this scenario corresponds to Region II. When the costs are relatively high ($c_F + 2c_V > \max\{\tilde{c}, \hat{c}\}$), this scenario corresponds to Region III. When $\hat{c} < \tilde{c}$, the regions are depicted in the left panel, while the case when $\hat{c} > \tilde{c}$ is illustrated in the right panel.

As shown in Figure 5, the results from Propositions 1 and 2 reveal a complementarity between the pro-price-competitive policy and compute subsidies: When one is ineffective, the other is effective. Specifically, in Region 1, where the costs are relatively low ($c_F + 2c_V < \min\{\tilde{c}, \hat{c}\}$), the pro-price-competitive policy does not improve consumer surplus, whereas compute subsidies are cost-effective in doing so. Conversely, in Region III, where the costs are relatively high, compute subsidies are cost-ineffective, while the pro-price-competitive policy becomes effective.

Our research provides important policy implications for policymakers seeking to enhance consumer surplus within the AI supply chain by offering structured policy recommendations under different cost conditions. As illustrated by the green lines in Region I of Figure 5, where compute and data preprocessing costs are relatively low ($0 < c_F + 2c_V < \min\{\hat{c}, \tilde{c}\}$), policymakers should focus solely on promoting competition in downstream markets, whether in terms of quality or price. In this case, as indicated by the red line in Region I, compute subsidies are not cost-effective because firms are already engaging in aggressive fine-tuning, leading to excessive public expenditure on subsidies with limited additional benefit to consumer surplus. When these costs are relatively high ($c_F + 2c_V > \max\{\hat{c}, \tilde{c}\}$), as illustrated by the green lines in Region III of Figure 5, policymakers should instead consider compute subsidies and pro-quality-competition policies. In this case, promoting price competition would reduce consumer surplus, as it discourages firms from investing in data for fine-tuning, thus lowering product quality. When these costs fall within an intermediate range ($\min\{\hat{c}, \tilde{c}\} < c_F + 2c_V < \max\{\hat{c}, \tilde{c}\}$), policymakers should consider adopting pro-quality-competitive

policy if $\hat{c} < \tilde{c}$, as illustrated by the green line in Region II of the left panel; otherwise ($\hat{c} > \tilde{c}$), they should consider all three policies, as illustrated by the green lines in Region II of the right panel

6. Effect of Different Regulatory Policies on Profits

In addition to consumer surplus, it is crucial to understand how pro-competitive policies and compute subsidies impact the profitability of both downstream firms and foundation model providers, as the success of any regulatory policy ultimately hinges on its acceptance by industry participants. A policy that not only enhances consumer surplus but also improves the profitability of both providers and downstream firms is more likely to gain traction, creating a "win–win–win" scenario where all stakeholders benefit. In this section, we begin by conducting a comparative statics analysis to assess how policies that promote competition in downstream markets impact the profits of both the foundation model provider and downstream firms, followed by an analysis of the effects of compute subsidies.

Understanding the effect of downstream competition on the profitability of both upstream and downstream firms remains a central concern in the study of industrial organization (Basak et al. 2023). The conventional view holds that increased competition among downstream firms reduces their profitability while enhancing the profits of upstream firms (Basak et al. 2023, Tyagi 1999, Bhattacharya et al. 2022). This perspective is based on the concept of double marginalization, which suggests that heightened competition among downstream firms weakens their pricing power, resulting in lower product prices (i.e., reduced double marginalization) and, in turn, higher demand for the product (Tyagi 1999). As a result, upstream firms are expected to earn higher profits from the expanded quantity sold to downstream firms, while downstream firms experience shrinking margins and reduced profitability. However, our analysis challenges this conventional wisdom, revealing that these typical outcomes do not necessarily hold in our setting. We summarize our findings in the following proposition and corollary.

PROPOSITION 3 **(Impact of Pro-competitive Policies on the Foundation Model Provider's Profits)**.

a. As price competition intensifies, the profits of the foundation model provider decrease if $\theta_q > \tilde{\theta}_q$, and increase otherwise.

b. As quality competition intensifies, the profits of the foundation model provider always increase.

For brevity, the expression for $\tilde{\theta}_q$ is provided in the Online Appendix.

This result highlights that intensified price competition in the downstream market can, under certain conditions, reduce the profits of the foundation model provider. To better understand how downstream price competition affects the provider's profitability, we apply the chain rule to analyze $\frac{d\pi_M}{d\theta_p}\big|_{p_I^*, p_F^*}$ and present the simplified expression below, with the full derivation available in the Online Appendix:

$$\left.\frac{d\pi_M}{d\theta_p}\right|_{p_I^*, p_F^*} = \sum_{i=1}^{2}\left(\frac{\partial \pi_M}{\partial p_i}\frac{\partial p_i(p_I, p_F)}{\partial \theta_p} + \frac{\partial \pi_M}{\partial V_i}\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\right)\Bigg|_{p_I^*, p_F^*}, \tag{9}$$

where the expression for π_M is provided in Equation (7), with $\frac{\partial \pi_M}{\partial p_i}\big|_{p_I^*, p_F^*} = -p_I^*$ and $\frac{\partial \pi_M}{\partial V_i}\big|_{p_I^*, p_F^*} = p_I^* + p_F^* - 2c_F V_i^*$.

Equation (9) demonstrates that intensified price competition influences the provider's profits by shaping downstream firms' decisions regarding both pricing and data usage for finetuning. On the one hand, consistent with conventional wisdom (i.e., reduced double marginalization), increased price competition prompts firms to lower their product prices ($\frac{\partial p_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$), which stimulates product demand and, in turn, increases the provider's inference revenue ($\frac{\partial \pi_M}{\partial p_i}\frac{\partial p_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} = -p_I^*\frac{\partial p_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} > 0$). On the other hand, intensified price competition encourages downstream firms to reduce their data usage for fine-tuning ($\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$), as discussed following Lemma 1. This decline in fine-tuning demand not only reduces fine-tuning revenue ($p_F^*\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$), but also degrades product quality, which in turn reduces inference revenue ($p_I^*\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$). Additionally, the reduced data usage lowers the compute costs associated with fine-tuning ($2c_F V_i^*\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$). However, despite the savings in compute costs, the overall effect of reduced data usage is a net decline in the provider's profits ($\frac{\partial \pi_M}{\partial V_i}\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$). The magnitude of this effect is influenced by the intensity of quality competition (i.e., θ_q). Specifically, as explained below, a larger θ_q strengthens the effect of reduced data volume.

When θ_q increases, the provider is incentivized to raise the fine-tuning price, as shown in Lemma 3. The resulting increase in p_F^* amplifies the loss of fine-tuning revenue caused by the decline in data volume resulting from intensified price competition ($p_F^*\frac{\partial V_i(p_I, p_F)}{\partial \theta_p}\big|_{p_I^*, p_F^*} < 0$). Consequently, the effect of reduced data volume becomes more pronounced as θ_q increases. Thus, when θ_q is relatively high (i.e, $\theta_q > \tilde{\theta}_q$), the adverse effect of reduced data volume becomes dominant and outweighs the positive effect of increased demand, leading to a decline in the provider's profits. This outcome, where the provider's profits decreases in response to intensified downstream price competition within the context of the AI supply chain, contrasts with established insights from traditional supply chains involving manufacturers and retailers (Tyagi 1999, Yenipazarli 2021). The key distinction lies in a novel mechanism we identify within the AI supply chain: Intensified price competition leads to reduced data usage for fine-tuning, which in turn diminishes both fine-tuning and inference revenues for the provider.

In contrast to the effects of price competition, Proposition 3b indicates that intensified quality competition always benefits the foundation model provider. To analyze this effect, we also apply the chain rule to analyze $\frac{d\pi_M}{d\theta_q}\big|_{p_I^*, p_F^*}$ and present the simplified expression below, with the full derivation available in the Online Appendix):

$$\left.\frac{d\pi_M}{d\theta_q}\right|_{p_I^*, p_F^*} = \left.\sum_{i=1}^{2}\left(\frac{\partial \pi_M}{\partial V_i}\frac{\partial V_i(p_I, p_F)}{\partial \theta_q} + \frac{\partial \pi_M}{\partial p_i}\frac{\partial p_i(p_I, p_F)}{\partial \theta_q}\right)\right|_{p_I^*, p_F^*},$$

where $\frac{\partial \pi_M}{\partial V_i}\big|_{p_I^*, p_F^*} = p_I^* + p_F^* - 2c_F V_i^*$ and $\frac{\partial \pi_M}{\partial p_i}\big|_{p_I^*, p_F^*} = -p_I^*$. As quality competition intensifies, firms increase their data usage for fine-tuning to enhance product quality ($\frac{\partial V_i(p_I, p_F)}{\partial \theta_q}\big|_{p_I^*, p_F^*} > 0$). This rise in fine-tuning

demand boosts fine-tuning revenue ($p_F^* \frac{\partial V_i(p_I,p_F)}{\partial \theta_q}\big|_{p_I^*,p_F^*} > 0$) and also increases inference demand due to the product quality improvements, leading to higher inference revenue ($p_I^* \frac{\partial V_i(p_I,p_F)}{\partial \theta_q}\big|_{p_I^*,p_F^*} > 0$). Although the increase in data volume raises the provider's compute costs, the net effect on profits remains positive because $((p_I^* + p_F^* - 2c_F V_i^*) \frac{\partial V_i(p_I,p_F)}{\partial \theta_q}\big|_{p_I^*,p_F^*} > 0)$. Additionally, as product quality improves, firms are incentivized to raise product prices, leading to increased double marginalization $\frac{\partial p_i(p_I,p_F)}{\partial \theta_q}\big|_{p_I^*,p_F^*} > 0)$. This price increase can reduce inference demand and thereby lower the provider's profits ($-p_I^* \frac{\partial p_i(p_I,p_F)}{\partial \theta_q}\big|_{p_I^*,p_F^*} < 0$). However, the positive impact on profitability from increased data volume always outweighs the negative impact from higher prices. Consequently, the provider's profits always increase as quality competition intensifies in the downstream market.

Next, we examine how increases in the intensity of price and quality competition, represented by higher values of θ_p and θ_q, respectively, affect downstream firms' profits and summarize our findings in the following corollary.

COROLLARY 2 **(Impact of Downstream Competition on Firms' Profits)**.

a. Intensified price competition may lead to an increase in downstream firms' profits.

b. Intensified quality competition always results in a decrease in downstream firms' profits.

Corollary 2 reveals that intensified price competition in the downstream market can actually increase downstream firms' profits. This finding challenges the conventional view that greater price competition always leads to reduced profitability (Tyagi 1999). The intuition behind this result is as follows. On the one hand, consistent with conventional wisdom, intensified price competition prompts firms to lower their product prices, which typically reduce revenues. On the other hand, in response to increased price competition, the upstream provider may lower both fine-tuning and inference prices, as shown in Lemmas 3 and 4. These price reductions decrease downstream firms' costs of using the provider's services. When the savings from these cost reductions exceed the revenue losses from lower product prices, downstream firms can see a net increase in profits.

Corollary 2 further indicates that intensified quality competition always reduces downstream firms' profits, which aligns with conventional expectations. As quality competition intensifies, firms must invest more in product quality to stay competitive, resulting in higher fine-tuning costs. Additionally, as established in Lemmas 3 and 4, the foundation model provider responds by increasing both fine-tuning and potentially inference prices, which further diminishes downstream profitability.

Proposition 3 and Corollary 2 offer key insights for foundation model providers, downstream firms, and policymakers. First, these results help both foundation model providers and downstream firms anticipate how their profits may evolve when pro-competitive policies are implemented in the downstream market. Such projections, given the resulting shifts in competitive dynamics, provide valuable guidance for strategic

planning (Rami Ali 2025). Second, these findings assist policymakers by clarifying how different pro-competitive interventions affect industry participants, thereby supporting more informed decision-making in AI supply chain regulation. In particular, policymakers should be aware of the potential downsides of pro-quality-competitive policies. While Proposition 1 demonstrates that such policies are robust in improving consumer surplus, Corollary 2 reveals that they simultaneously reduce profitability for downstream firms, complicating their practical implementation. In contrast, Proposition 3 and Corollary 2 suggest that pro-price-competitive policies may benefit both foundation model providers and downstream firms. To assess whether such policies can deliver a "win–win–win" outcome that benefits consumers, downstream firms, and foundation model providers, we examine this possibility in the following corollary.

COROLLARY 3 **(Win–Win–Win Outcome of Pro-price-competitive Policy)**. *There exist cases in which a pro-price-competitive policy increases consumer surplus while simultaneously boosting the profits of both the foundation model provider and downstream firms.*

Corollary 3 indicates that a pro-price-competitive policy can generate a "win–win–win" outcome, benefiting all stakeholders. In contrast, a pro-quality-competitive policy typically increases consumer surplus at the cost of reduced firm profitability. This highlights the potential advantage of prioritizing price competition over quality competition, suggesting that policymakers should weigh the broader benefits of pro-price-competitive policies, even though pro-quality-competitive policies are always effective in enhancing consumer surplus.

In addition to pro-competitive policies, we now examine another key policy instrument—compute subsidies—to explore how subsidizing access to computing power affects the profitability of both the foundation model provider and downstream firms. The main result is summarized in Proposition 4.

PROPOSITION 4 **(Impact of Compute Subsidies on Profits)**.

a. Downstream firms' profits decrease due to compute subsidies only when $c_F > \hat{c}_F$.

b. The foundation model provider's profits always increase When compute subsidies are provided.

For brevity, the expression for $\hat{c}_F$ is provided in the Online Appendix.

Proposition 4 highlights that while compute subsidies reduce compute costs for providers, they can unexpectedly decrease downstream firms' profits under certain conditions. This counter-intuitive outcome arises from opposing effects. On the one hand, compute subsidies lower the fine-tuning price, as discussed following Lemma 5. This price reduction decreases the fine-tuning fees paid by downstream firms, which could potentially boost their profits. Additionally, the lower fine-tuning price encourages firms to use more data for fine-tuning, thereby improving product quality. The resulting quality enhancement boosts product demand and, in turn, increases downstream firms' revenues. On the other hand, the increased data volume also generates two negative consequences for downstream firms. First, it directly raises their data preprocessing

costs. Second, the improved product quality drives up inference demand, prompting the provider to raise the inference price, and consequently, the inference fees downstream firms must pay. Downstream firms' profits decline when the combined increase in data preprocessing costs and inference fees exceeds the savings from the reduced fine-tuning fees. This outcome is particularly likely when the coefficient of compute costs (i.e., c_F) is relatively high. A high c_F results in a high fine-tuning price, which in turn leads to limited data volume. As a result, the savings in fine-tuning fees induced by the lower fine-tuning price due to compute subsidies are minimal, and the positive effect of the subsidy is outweighed by the increased data preprocessing costs and inference fees, ultimately reducing downstream firms' profits.

In contrast to downstream firms' profits, the foundation model provider's profits always increase when compute subsidies are introduced. This is because subsidies directly lower the compute costs for the fine-tuning service. Moreover, they drive a reduction in fine-tuning prices, encouraging downstream firms to invest more in product quality through fine-tuning. As product quality improves, the demand for inference services rises. Overall, the combination of reduced compute costs and increased inference revenue enhances the foundation model provider's profits. Since Proposition 4 indicates that subsidies can enhance the profitability of both the provider and downstream firms, this regulatory instrument may lead to a "win-win-win" outcome. We summarize the condition for this outcome in the following corollary.

COROLLARY 4 **(Win–Win–Win Outcome of Compute Subsidies).** *If $c_F < \hat{c}_F$, the provision of compute subsidies boosts consumer surplus and increases the profits of both the foundation model provider and downstream firms.*

Corollary 4 indicates that compute subsidies can create a "win–win–win" outcome, benefiting all stakeholders when compute costs are relatively low. The mechanism behind this outcome has been discussed following Lemma 5 and Proposition 4. Specifically, compute subsidies reduce the provider's compute costs, prompting the provider to lower the fine-tuning price. This reduction, in turn, lowers downstream firms' costs and strengthens their incentives to co-create and improve quality, ultimately benefiting consumers.

Our findings have important policy implications for the AI supply chain. Corollary 4 and Proposition 2 suggest that policymakers should consider implementing compute subsidies when compute costs are relatively low (i.e., $c_F < \min\{\hat{c}_F, \hat{c} - 2c_V\}$). In such cases, compute subsidies not only effectively enhance consumer surplus but also simultaneously improve the profitability of both providers and downstream firms, making them a desirable policy intervention.

Beyond this specific recommendation, synthesizing the findings from Sections 5 and 6 reveals a fundamental trade-off in policy design between the robustness of consumer benefits and the feasibility of achieving industry-wide gains, highlighting the need for carefully calibrated interventions. We identify two policy classes with distinct strengths and limitations: First, pro-quality-competitive policies are guaranteed to enhance consumer surplus. However, they cannot achieve a "win-win-win" outcome due to their adverse

impact on the downstream firms' profitability. Second, Pro-price-competitive policies and compute subsidies represent a different pattern. While they are not guaranteed to enhance consumer surplus, they have the potential to create a "win-win-win" outcome, benefiting all stakeholders. This potential for broad industry buy-in makes the second class of policies more practical to implement, even if their benefits to consumers are conditional. From this perspective, policymakers face a trade-off: policies that robustly help consumers may fail to benefit participants across the AI supply chain, whereas policies that can benefit both upstream and downstream participants of AI supply chain may deliver consumer benefits only contingently.

7. Effect of Declining Compute Costs on Profits and Consumer Surplus

Beyond assessing the immediate impacts of regulatory policies, it is crucial to account for how their adoption may need to evolve over time in response to the rapidly changing landscape of AI markets. A key factor driving this shift is the ongoing reduction in compute costs, largely due to advancements in graphics processing units (GPUs), which are central to fine-tuning AI models (NVIDIA 2024, Diehl 2025). To better understand these dynamics, we conduct a comparative statics analysis to explore how declining compute costs influence consumer surplus, as well as the profits of both the foundation model provider and downstream firms. We also examine how regulatory policies should adapt to ensure that consumer-surplus-enhancing outcomes are maintained.

Intuitively, lower compute costs should enhance profitability across the AI ecosystem and provide benefits to consumers (Osman 2024). However, prior research suggests that reductions in IT costs can reshape competitive dynamics and may even erode firms' profits in the setting with only downstream firms (Demirhan et al. 2005). Because our setting features an AI supply chain—with foundation model providers, downstream firms, and consumers—the competitive dynamics differ from those studied in the literature. It is therefore unclear how declining compute costs will affect profits at each layer and whether the impacts on providers and downstream firms will diverge. Moreover, in the AI supply chain, where downstream firms co-create value through fine-tuning, these effects could impact the firms' incentives to invest in quality, which in turn affect consumer surplus. This highlights the importance of carefully examining how changes in compute costs affect consumer surplus as well as the profitability of both providers and downstream firms within the AI supply chain. The findings of this analysis are presented in Proposition 5.

PROPOSITION 5 **(Impact of Compute Costs)**. *As the coefficient of compute costs (i.e., c_F) declines,*

a. downstream firms' profits decrease if $c_F > \hat{c}_F$ and increase otherwise;

b. the foundation model provider's profits and consumer surplus always increase.

For brevity, the expression for $\hat{c}_F$ is provided in the Online Appendix.

Contrary to intuition, our results reveal that a decrease in c_F may result in a decline in firms' profits. This counterintuitive effect arises from the following mechanism. Using the chain rule, we analyze $\frac{d\pi_i}{dc_F}\big|_{p_I^*, p_F^*}$ as follows:

$$\left.\frac{d\pi_i}{dc_F}\right|_{p_I^*, p_F^*} = \left.\left(\frac{\partial \pi_i}{\partial p_F}\frac{dp_F^*}{dc_F} + \frac{\partial \pi_i}{\partial p_I}\frac{dp_I^*}{dc_F} + \frac{\partial \pi_i}{\partial V_i}\frac{dV_i^*}{dc_F} + \frac{\partial \pi_i}{\partial p_i}\frac{dp_i^*}{dc_F}\right)\right|_{p_I^*, p_F^*},$$

where the expression for π_i is given in Equation (6).

As c_F decreases, two opposing forces shape its impact on downstream firms' profits. On the one hand, a lower c_F reduces the foundational model provider's marginal cost of fine-tuning, leading to a decrease in the fine-tuning price ($-\frac{dp_F^*}{dc_F} < 0$). This reduction lowers the firms' fine-tuning expenses, thereby increasing their profits ($-\big(\frac{\partial \pi_i}{\partial p_F}\frac{dp_F^*}{dc_F}\big)\big|_{p_I^*, p_F^*} > 0$). On the other hand, lower compute costs also affect profits through changes in the inference price, firms' product prices, and the data volume used for fine tuning. The combined effect of these three channels is negative, i.e., $-\big(\frac{\partial \pi_i}{\partial p_I}\frac{dp_I^*}{dc_F} + \frac{\partial \pi_i}{\partial V_i}\frac{dV_i^*}{dc_F} + \frac{\partial \pi_i}{\partial p_i}\frac{dp_i^*}{dc_F}\big)\big|_{p_I^*, p_F^*} < 0$. This occurs because a lower fine-tuning price intensifies quality competition, which prompts firms to increase their investment in quality, which raises their expenditures (i.e., data preprocessing costs and fine-tuning fees) and directly reduces their profits. Furthermore, the improvement in product quality increases consumer demand, which, in turn, raises inference demand for the provider. In response, the provider may raise the inference price charged to firms, further eroding their profits. Together, these negative effects, combined with the positive effect from the lower fine-tuning price, generate a non-monotonic relationship between c_F and downstream firms' profitability. When c_F is relatively high, the high compute cost of fine-tuning induces the provider to set a correspondingly high fine-tuning price, which ultimately leads to a low data volume. Given this low data volume, the reduction in the fine-tuning price resulting from declining compute costs leads to only a small profit gain for firms ($\frac{\partial \pi_i}{\partial p_F}\big|_{p_I^*, p_F^*} = -V_i^*$), yielding a weak positive effect. Consequently, the negative effects dominate, causing downstream firms' profits to decrease as compute costs decline.

These findings contribute to the literature on the implications of declining IT costs in competitive markets. Early work by Demirhan et al. (2005) shows that declining IT costs generally benefit firms by enabling greater investments, which in turn enhance product quality and increase profits. However, Demirhan et al. (2007) find that when switching costs between competing firms' products are sufficiently high, declining IT costs can reduce firm profitability. This occurs because lower IT costs prompt firms to engage in aggressive quality competition, which ultimately erodes their profits. This result is consistent with our finding that firms may also be worse off under declining compute costs. However, our model introduces a different mechanism behind this outcome. In Demirhan et al. (2007), overinvestment arises from intertemporal competition, where high switching costs lock consumers into future purchases, incentivizing firms to overinvest in the initial period to secure long-term demand. In contrast, in our model, the pricing behavior of the foundation model provider plays a central role. Specifically, as compute costs fall, the provider lowers the fine-tuning price,

encouraging firms to increase their investment in product quality. This mechanism reflects the co-creation dynamics of the AI supply chain, which, to our knowledge, has not been identified in the existing literature.

Proposition 5b further shows that the foundation model provider's profits always increase as c_F decreases. This occurs because lower compute costs (c_F) reduce the provider's expenses for delivering fine-tuning services, which in turn leads to a reduction in the fine-tuning price. The lower price incentivizes downstream firms to expand data volumes and enhance product quality, which boosts inference demand and increases inference revenue. Taken together, the reduction in fine-tuning costs and the increase in inference revenue result in higher provider profits as compute costs decline. Similarly, consumer surplus increases as compute costs decline. This is because lower compute costs lead to a reduction in the provider's fine-tuning price, which enhances downstream firms' incentives to engage in the co-creation process and improve product quality. The resulting improvements in product quality ultimately increase consumer surplus.

While Proposition 5 demonstrates that consumer surplus increases as compute costs decline, it remains unclear whether regulatory interventions that are effective under current cost structures will continue to be appropriate as costs decrease further. To address this, we examine how declining compute costs affect the conditions for effective policy, as outlined in Proposition 1 and Proposition 2. The results are summarized in the following corollary.

COROLLARY 5 (**Impact of Declining Compute Costs on Policies**). *As compute costs decline, (i) compute subsidies are more likely to be effective in enhancing consumer surplus, (ii) the pro-price-competitive policy becomes less likely to be effective in enhancing consumer surplus, and (iii) The pro-quality-competitive policy remains consistently effective.*

Recall that compute subsidies are effective only when compute and data preprocessing costs are relatively low, while the pro-price-competitive policy is effective only when these costs are relatively high. Thus, Corollary 5 suggests that as compute costs decline, compute subsidies may shift from being ineffective to effective. This implies that policymakers should consider adopting this policy in the future, even if it appears ineffective under current cost conditions. In contrast, as compute costs continue to fall, the pro-price-competitive policy may shift from being an effective tool for enhancing consumer surplus to one that harms it. Therefore, policymakers should exercise caution with this policy and consider phasing it out once compute costs fall to a relatively low level. Finally, the declining trend in compute costs does not significantly affect the effectiveness of the pro-quality-competitive policy. This policy consistently increases consumer surplus, regardless of compute cost levels, making it a robust tool for enhancing consumer surplus. Overall, since compute costs are expected to continue declining, our findings highlight the importance of ongoing monitoring and adaptive policymaking.

8. Discussions and Conclusions

The rapid innovation in AI has brought regulatory concerns to the forefront, driven by fears that firms may exploit AI technologies to consolidate market power and capture an excessive share of consumer surplus. This study analyzes these regulatory concerns within the context of the emerging AI supply chain, a systemic perspective that encompasses the interactions between upstream and downstream AI markets. We analyze two types of policies within this supply chain and evaluate their effectiveness. To protect consumers, pro-competitive policy are commonly used; however, the specific approach to promoting competition is critical. Promoting downstream quality competition consistently improves consumer surplus, while promoting downstream price competition is beneficial only when at least compute or data preprocessing costs are high. Additionally, we introduce a new class of policy aimed at enhancing surplus within the AI supply chain: compute subsidies. Our analysis shows that compute subsidies, which can effectively enhancing consumer surplus when both the compute and data preprocessing costs are low, complement pro-price-competitive policies. We also examine the impact of these policies on the participants of the AI supply chain. We find that while pro-quality-competitive policies tend to reduce downstream firms' profits, they increase the profits of the foundation model provider. In contrast, both pro-price-competitive policy and compute subsidies can increase consumer surplus as well as the profits of both providers and downstream firms, leading to a "win-win-win" outcome in which all stakeholders benefit. Finally, we explore the implications of declining compute costs. As these costs continue to fall, pro-price-competitive policies may shift from being an effective tool for enhancing consumer surplus to a harmful one that diminishes it. In contrast, compute subsidies may transition from an ineffective policy to an effective one as compute costs decline.

Our study makes significant theoretical contributions. Specifically, we take a crucial first step in characterizing regulation within the AI supply chain by systematically analyzing the interactions between upstream and downstream participants. This is particularly important given the limited theoretical research on regulation in traditional supply chain contexts. Furthermore, our study highlights the unique characteristics of the AI supply chain that distinguish it from existing studies on traditional supply chains and software markets (the latter being relevant because fine-tuning and inference of foundation models are delivered as software-enabled services). First, unlike traditional supply chains, the development of fine-tuned models involves co-creation between foundation model providers and downstream firms. Additionally, the provider generates revenue through two streams: inference and fine-tuning, whereas manufacturers in traditional supply chain typically rely on wholesale revenue. Second, the co-creation process in the AI supply chain differs from other well-known co-creation processes among firms. In traditional settings, co-creation typically takes place independently across participants, with each party able to exert effort autonomously to enhance the quality of final product, such as those involving vendor–client collaborations for enterprise systems (e.g., ERP); cases in the computer industry where one firm develops the central processor while another develops the operating

system; and settings in the smartphone industry where one firm designs the device and operating platform while other firms develop applications for it (Demirezen et al. 2020, Puyang et al. 2025, Gupta et al. 2023, Ceccagnoli et al. 2012). In contrast, the fine-tuning of foundation models involves a mutually dependent form of co-creation. The foundation model provider not only supplies the base model but also offers the computational infrastructure required for downstream firms to use their domain-specific data to improve model quality. When downstream firms contribute more data to enhance model performance, this increases the compute demand and thus the provider's compute costs. Such interdependence in the co-creation process is largely absent from traditional co-creation settings. Third, the customization process in the AI supply chain differs significantly from that in traditional software markets. In conventional settings, software providers sell licenses to users, who then customize and deploy the software on their own infrastructure, usually for a one-time fixed fee (Ma and Seidmann 2015, Zhang et al. 2020). In contrast, fine-tuned AI models are typically hosted on the provider's cloud infrastructure and offered through a pay-as-you-go pricing model. Fourth, this deployment mechanism and pricing structure place both fine-tuning and inference services within the broader category of SaaS. However, these services differ significantly from traditional SaaS in terms of customization flexibility. While traditional SaaS applications often offer limited customization options (CompTIA 2025, Zhang et al. 2020), fine-tuning services enable users to adapt AI models with much greater flexibility to meet specialized needs.

In the AI supply chain, we also identify several novel mechanisms that arise from its unique co-creation structure. First, pro–price-competitive policies in downstream markets increase the provider's incentive to engage in co-creation and induce the provider to lower the fine-tuning price. This, in turn, improves product quality and enhances consumer surplus. In contrast, pro–quality-competitive policies weaken the provider's co-creation incentive and lead the provider to raise the fine-tuning price, which negatively affects the product quality. Second, while pro–price-competitive policies reduce product prices and are therefore expected—by conventional wisdom—to increase the provider's profits through reduced double marginalization, they also weaken downstream firms' incentives for quality creation. This reduces the volume of domain-specific data used in fine-tuning, thereby lowering fine-tuning and inference revenue and ultimately decreasing the provider's profits. Third, as compute costs decline or compute subsidies are introduced, the provider gains stronger incentives to participate in co-creation. As a result, the provider reduces the fine-tuning price, encouraging downstream firms to invest more in product quality. This mechanism reflects a unique form of co-creation dynamics within the AI supply chain—one that is absent from existing studies on the implications of declining compute costs (Demirhan et al. 2005, 2007).

From a practical standpoint, our research provides valuable insights for policymakers seeking to enhance consumer surplus within the AI supply chain. First, we show that pro-price-competitive policies and compute subsidies are not universally effective; rather, their effectiveness depends on the underlying cost conditions (as shown in Proposition 1 and Proposition 2). This enables policymakers to identify which intervention is

better suited for specific market conditions. Importantly, these two policies can function as complementary regulatory instruments, as each tends to be effective when the other is not. For downstream markets that handle complex, sensitive, and high-stakes data and incur high data preprocessing costs—such as healthcare and finance (Reddit 2025)—pro-price-competitive policies may be ineffective or even detrimental to consumer welfare, while compute subsidies are likely to yield substantial benefits. Conversely, in markets with relatively low data preprocessing costs, such as e-commerce, pro-price-competitive policies can effectively enhance consumer surplus, whereas compute subsidies may have limited impact. Second, many governments already provide compute subsidies at varying rates—for instance, Beijing offers subsidies covering up to 30% of compute expenses, while Shenzhen covers up to 50%. Our results show that excessively high subsidy rates can become counterproductive: their effectiveness diminishes and may even turn negative once the subsidy rate exceeds a certain threshold (as shown in Corollary 1). Accordingly, a moderate subsidy rate maximizes net gains in consumer surplus and provides clear guidance for effective subsidy design. Third, we find that pro-price-competitive policies and compute subsidies are generally more viable and more likely to gain support from participants across the AI supply chain than pro-quality-competitive policies (as shown in Proposition 3). This is because both approaches have the potential to generate a "win–win–win" outcome across consumers, downstream firms, and foundation model providers. Although pro-quality-competitive policies always improve consumer welfare, they consistently reduce downstream firms' profits and thus cannot deliver such a mutually beneficial outcome—an important consideration for regulatory adoption. Fourth, our analysis underscores the importance of dynamic and adaptive policymaking. As GPUs improve and compute costs decline, the effectiveness of policy tools may reverse (as shown in Corollary 5): pro-price-competitive policies may shift from being effective to ineffective, while compute subsidies may shift from ineffective to effective. Policymakers should therefore closely monitor cost trends and adjust regulatory strategies accordingly, using our framework as a guide for navigating evolving technological conditions. Finally, as compute costs continue to fall, our results suggest that policymakers should place greater emphasis on compute subsidies (as shown in Corollary 4 and Proposition 2). In future low-cost environments, compute subsidies not only effectively increase consumer surplus but also help achieve a robust "win–win–win" outcome, making them an especially promising lever for fostering socially beneficial AI ecosystems.

Acknowledgments

The authors thank participants at the Workshop on Information Systems and Economics (Hyderabad, 2023), the 34th Annual POMS Conference (Minneapolis, 2024), and the 3rd Annual Business & Generative AI Conference (San Francisco, 2025), as well as seminar participants at the University of Alberta, for their valuable comments and insightful discussions. All authors certify that they have no affiliations with, or involvement in, any organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Acemoglu D (2021) Harms of AI. Working paper, National Bureau of Economic Research.

AI Now Institute (2023) Chatgpt and more: Large scale ai models entrench big tech power. Accessed August 11, 2023, https://ainowinstitute.org/publication/large-scale-ai-models#footnote-list-2.

Anderson A, Hong HG, Jacobs E (2022) Are e-bike subsidies cost effective in mitigating carbon emissions? Working paper, National Bureau of Economic Research.

Armstrong M, Chen Y (2009) Inattentive consumers and product quality. *Journal of the European Economic Association* 7(2-3):411–422.

Assad S, Clark R, Ershov D, Xu L (2024) Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. *Journal of Political Economy* 132(3):723–771.

Austrian Federal Competition Authority (2024) Artificial intelligence, data and competition – note by austria. Technical report, Organisation for Economic Co-operation and Development.

Banker RD, Khosla I, Sinha KK (1998) Quality and competition. *Management science* 44(9):1179–1192.

Basak D, Marjit S, Mukherjee A (2023) Competition and profit in a supply chain. Preprint, submitted August 30, http://dx.doi.org/10.2139/ssrn.4557289.

Bhati D (2024) Google AI boss says company is investing more than $100 billion in AI to be ahead of its competitors. Accessed November 30, 2024, https://www.indiatoday.in/technology/news/story/google-ai-boss-says-company-is-investing-more-than-100-billion-in-ai-to-be-ahead-of-its-competitors-2528722-2024-04-18.

Bhattacharya S, Gong J, Wattal S (2022) Competitive poaching in search advertising: Two randomized field experiments. *Information Systems Research* 33(2):599–619.

Boggs J (2025) New ftc rule forces ticketmaster to display all prices upfront. Accessed Oct 12, 2025, https://www.scrippsnews.com/life/money/new-ftc-rule-forces-ticketmaster-to-display-all-prices-upfront.

Bohlmann JD, Golder PN, Mitra D (2002) Deconstructing the pioneer's advantage: Examining vintage effects and consumer valuations of quality and variety. *Management Science* 48(9):1175–1195.

Cajueiro DO, Celestino VRR (2025) A comprehensive review of artificial intelligence regulation: Weighing ethical principles and innovation. *Journal of Economy and Technology* 4:77–91.

CCPC (2025) Digital content / services. Accessed Oct 9, 2025, https://www.ccpc.ie/business/help-for-business/guidelines-for-business/digital-content-services/.

Ceccagnoli M, Forman C, Huang P, Wu D (2012) Cocreation of value in a platform ecosystem! the case of enterprise software. *MIS quarterly* 263–290.

Cen SH, Hopkins A, Ilyas A (2024a) AI supply chains (and why they matter). Accessed November 7, 2024, https://aipolicy.substack.com/p/supply-chains-2.

Cen SH, Hopkins A, Struckman I, Videgaray L (2024b) Three proposals for regulating AI. Accessed November 7, 2024, https://aipolicy.substack.com/p/supply-chains-5.

Competition and Markets Authority (2023) AI Foundation Models: Initial Report. Accessed June 21, 2025, https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

Competition Bureau (2024) Artificial intelligence and competition. Accessed August 30, 2024, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/artificial-intelligence-and-competition.

Competition Bureau Canada (2025) Report on summit 2024: Competition in the ai era. Accessed Oct 12, 2025, https://competition-bureau.canada.ca/en/how-we-foster-competition/education-and-outreach/report-summit-2024-competition-ai-era.

CompTIA (2025) What is SaaS? Accessed May 15, 2025, https://www.comptia.org/content/articles/what-is-saas.

Demirezen EM, Kumar S, Shetty B (2020) Two is better than one: A dynamic analysis of value co-creation. *Production and operations management* 29(9):2057–2076.

Demirhan D, Jacob VS, Raghunathan S (2005) Information technology investment strategies under declining technology cost. *Journal of Management Information Systems* 22(3):321–350.

Demirhan D, Jacob VS, Raghunathan S (2007) Strategic it investments: The impact of switching cost and declining it cost. *Management Science* 53(2):208–226.

Design Gurus (2025) Is OpenAI a SaaS company? Accessed Feb 24, 2025, https://www.designgurus.io/answers/detail/is-openai-a-saas-company.

Diehl S (2025) Introduction to fine-tuning large language models. Accessed June 1, 2025, https://www.stephendiehl.com/posts/training_llms/?utm_source=chatgpt.com.

Dreyer C (2025) 15 best legal ai tools in 2025: Features, pricing & how to choose. Accessed Oct 12, 2025, https://rankings.io/blog/legal-ai-tools/.

Eldridge M (2025) The limitations of ai in legal operations: Why technology alone won't solve all legal challenges. Accessed Oct 12, 2025, https://www.acc.com/resource-library/limitations-ai-legal-operations-why-technology-alone-wont-solve-all-legal.

FCC (2025) Fcc requires broadband providers to display labels to help consumers. Accessed Oct 12, 2025, https://www.fcc.gov/document/fcc-requires-broadband-providers-display-labels-help-consumers.

Fisher S (2024) 10 top ai companies in law. Accessed November 30, 2024, https://www.vktr.com/ai-market/10-top-ai-companies-in-law/.

FTC (2025) Anticompetitive practices. Accessed May 30, 2025, https://www.ftc.gov/enforcement/anticompetitive-practices#.

FTC (2025a) Federal trade commission announces final rule banning fake reviews and testimonials. Accessed Oct 12, 2025, https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials.

FTC (2025b) Ftc announces crackdown on deceptive ai claims and schemes. Accessed May 30, 2025, https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes.

FTC (2025c) Ftc sends refunds to consumers affected by fashion nova's deceptive review practices. Accessed Oct 12, 2025, https://www.ftc.gov/news-events/news/press-releases/2025/01/ftc-sends-refunds-consumers-affected-fashion-novas-deceptive-review-practices.

FTC (2025d) The rule on unfair or deceptive fees: Frequently asked questions. Accessed Oct 12, 2025, https://www.ftc.gov/business-guidance/resources/rule-unfair-or-deceptive-fees-frequently-asked-questions.

Government of Canada (2024) Promotion and advocacy. Accessed November 7, 2024, https://competition-bureau.canada.ca/how-we-foster-competition/promotion-and-advocacy.

Guo P (2009) A survey of software as a service delivery paradigm. *TKK T-110.5190 Seminar on Internetworking*, 1–3.

Guo Z, Ma D (2018) A model of competition between perpetual software and software as a service. *MIS Quarterly* 42(1):101–120.

Gupta S, Roy A, Kumar S, Mudambi R (2023) When worse is better: Strategic choice of vendors with differentiated capabilities in a complex cocreation environment. *Management Science* 69(5):2833–2851.

Hrdy AS, Ceccio J, Zeytoomian C (2025) FTC Moves to Expand AI Deployment Oversight with Operation AI Comply. Accessed May 30, 2025, https://www.morganlewis.com/pubs/2024/10/ftc-moves-to-expand-ai-deployment-oversight-with-operation-ai-comply.

Huizenga E (2025) Fine-tuning gemini: Best practices for data, hyperparameters, and evaluation. Accessed May 16, 2025, https://medium.com/google-cloud/fine-tuning-gemini-best-practices-for-data-hyperparameters-and-evaluation-65f7c7b6b15f.

Insights A (2024) Training AI Models with OpenAI API: How to Handle Large Datasets and GPU Usage. Accessed November 30, 2024, https://medium.com/muthoni-wanyoike/training-ai-models-with-openai-api-how-to-handle-large-datasets-and-gpu-usage-9f1bbcfba38f.

ITSoli (2025) The future of custom llm fine-tuning: Trends to watch. Accessed Apr 29, 2025, https://itsoli.ai/the-future-of-custom-llm-fine-tuning-trends-to-watch/.

Jiang Z, Qu XS, Jain DC (2019) Optimal market entry timing for successive generations of technological innovations1. *MIS quarterly* 43(3):787–806.

Karanika V (2013) The cloud, mobility and beyond. Accessed January 28, 2023, https://www.cisco.com/c/dam/global/el_gr/assets/connect2013/pdfs/015_idc_evangelos_karanikas.pdf.

Kardach J (2024) Fine-tuning LLMs for Your Industry: Optimal Data Labeling Strategies. Accessed November 30, 2024, https://www.toptal.com/artificial-intelligence/llm-data-labeling.

Kim B, Chen P, Mukhopadhyay T (2009) An economic analysis of the software market with a risk-sharing mechanism. *International Journal of Electronic Commerce* 14(2):7–40, ISSN 1086-4415.

Leblanc T, Dailard S, Leavy K, Velez N (2025) FTC Warns to 'Keep Your AI Claims in Check' in New AI Guidance. Accessed May 30, 2025, https://cdp.cooley.com/ftc-warns-to-keep-your-ai-claims-in-check-in-new-ai-guidance/.

Ma D, Kauffman RJ (2014) Competition between software-as-a-service vendors. *IEEE Transactions on Engineering Management* 61(4):717–729.

Ma D, Seidmann A (2015) Analyzing software as a service with per-transaction charges. *Information Systems Research* 26(2):360–378.

Mahari R, Pentland A (2024) Regulation by design: A new paradigm for regulating ai systems. Preprint, submitted April 2, https://ssrn.com/abstract=4753029.

Matsushima N (2006) Industry profits and free entry in input markets. *Economics Letters* 93(3):329–336.

Menear H (2024) Large language models could drive the next industrial revolution. Accessed August 30, 2024, https://interface.media/blog/2024/05/07/large-language-models-could-drive-the-next-industrial-revolution/.

Mohammadi B, Malik N, Derdenger T, Srinivasan K (2024) Regulating explainable artificial intelligence (xai) may harm consumers. *Marketing Science* .

Muller J, Kruger J, Enderlein S, Helmich M, Zeier A (2009) Customizing enterprise software as a service applications: Back-end extension in a multi-tenancy environment. *Enterprise Information Systems: 11th International Conference, ICEIS 2009, Milan, Italy, May 6-10, 2009. Proceedings 11*, 66–77 (Springer).

Munappy AR, Bosch J, Olsson HH, Arpteg A, Brinne B (2022) Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software* 191:111359.

Narayanan D, Shoeybi M, Casper J, LeGresley P, et al. (2021) Efficient large-scale language model training on gpu clusters using megatron-lm. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–15.

Ngo V (2025) Challenges of product analytics in the era of generative ai. Accessed Oct 12, 2025, https://medium.com/@vuongngo/challenges-of-product-analytics-in-the-era-of-generative-ai-3abf42a44e76.

NVIDIA (2024) Artificial intelligence. Accessed November 7, 2024, https://www.nvidia.com/en-us/glossary/artificial-intelligence/.

OpenAI (2023) Fine-tuning. Accessed August 11, 2024, https://platform.openai.com/docs/guides/fine-tuning.

OpenAI (2024a) Harvey. Accessed November 30, 2024, https://openai.com/index/harvey/.

OpenAI (2024b) Pricing. Accessed August 22, 2024, https://openai.com/api/pricing/.

Osman J (2024) Can Nvidia's 'Hyper Moore's Law' Spark An AI Revolution? Forbes (November 7), https://www.forbes.com/sites/jimosman/2024/11/07/can-nvidias-hyper-moores-law-spark-an-ai-revolution/.

Precedence Research (2023) Artificial Intelligence (AI) Market Size. Accessed August 11, 2024, https://www.precedenceresearch.com/artificial-intelligence-market.

Puyang X, Gu Z, Chen RR, Li J (2025) Vertical competition on a common platform. *Marketing Science* .

Rami Ali (2025) Profitability forecasts: What they are and how to create them. Accessed Apr 25, 2025, https://www.netsuite.com/portal/resource/articles/financial-management/profitability-forecast.shtml.

Reddit (2025) What industry has the worst data? Accessed Nov 27, 2025, https://www.reddit.com/r/MachineLearning/comments/1eyj7vq/d_what_industry_has_the_worst_data/.

Reuters T (2024) Gpt-4 alone is not a reliable legal solution—but it does enable one. Accessed November 30, 2024, https://casetext.com/blog/cocounsel-harnesses-gpt-4s-power-to-deliver-results-that-legal-professionals-can-rely-on/.

Robinson B (2025) Consumer protection in the age of ai: The ftc's approach to ai regulation. Accessed Oct 9, 2025, https://iapp.org/news/a/consumer-protection-in-the-age-of-ai-the-ftcs-approach-to-ai-regulation.

Sala F (2024) Getting better performance from foundation models (with less data). Accessed November 30, 2024, https://snorkel.ai/blog/getting-better-performance-from-foundation-models-with-less-data/.

Sastry G, Heim L, Belfield H, Anderljung M, Brundage M, Hazell J, O'Keefe C, Hadfield GK, Ngo R, Pilz K, et al. (2024) Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797* .

Schneier B (2015) Should companies do most of their computing in the cloud? (part 1). Accessed June 28, 2025, https://www.schneier.com/blog/archives/2015/06/should_companie.html.

Schultz C (2009) Transparency and product variety. *Economics Letters* 102(3):165–168.

Shaver N (2025) The use of large language models in legaltech. Accessed Mar 26, 2025, https://www.legaltechnologyhub.com/contents/the-use-of-large-language-models-in-legaltech/.

Singh N, Vives X (1984) Price and quantity competition in a differentiated duopoly. *The Rand journal of economics* 546–554.

Singh S (2024) Beginner's Guide: Using Foundation Models in ML Projects. Accessed November 30, 2024, https://www.labellerr.com/blog/beginners-guide-using-foundation-models-in-ml-projects/.

Stryker C (2025) What is a pretrained model? Accessed July 20, 2025, https://www.ibm.com/think/topics/pretrained-model.

Sun C, Ji Y, Geng X (2023) Which enemy to dance with? a new role of software piracy in influencing antipiracy strategies. *Information Systems Research* 34(4):1711–1727.

Sun C, Shi Z, Liu X, Ghose A (2025) The effect of voice ai on digital commerce. *Information Systems Research* 36(2):1147–1166.

Sun W, Zhang X, Guo CJ, Sun P, Su H (2008) Software as a service: Configuration and customization perspectives. *2008 ieee congress on services part ii (services-2 2008)*, 18–25 (IEEE).

Tirole J (1988) *The theory of industrial organization* (MIT Press).

Tsay AA, Agrawal N (2000) Channel dynamics under price and service competition. *Manufacturing & Service Operations Management* 2(4):372–391.

Tucker C (2024) How does competition policy need to change in a world of artificial intelligence? *Oxford Review of Economic Policy* 40(4):834–842.

Tyagi RK (1999) On the effects of downstream entry. *Management science* 45(1):59–73.

Vohra DK (2024) 5 Real-world Applications of Large AI Models. Accessed November 30, 2024, https://www.hyperstack.cloud/blog/case-study/real-world-applications-of-large-ai-models.

Wang Q, Huang Y, Jasin S, Singh PV (2023) Algorithmic transparency with strategic users. *Management Science* 69(4):2297–2317.

Wang Y, Ha AY, Tong S (2022) Sharing manufacturer's demand information in a supply chain with price and service effort competition. *Manufacturing & Service Operations Management* 24(3):1698–1713.

Wang Z, Guo H, Liu D (2024) Economics of analytics services on a marketplace platform. *MIS Quarterly* 48(2).

Wikipedia (2023) GPT-4. Accessed August 10, 2023, https://en.wikipedia.org/wiki/GPT-4#cite_note-30.

Wu D (2024) Chinese cities offer subsidies to boost access to the computing power needed for AI. South China Morning Post (December 27), https://www.scmp.com/economy/china-economy/article/3292478/chinese-cities-offer-subsidies-boost-access-computing-power-needed-ai.

Wu Y, Zhu F (2022) Competition, contracts, and creativity: Evidence from novel writing in a platform market. *Management Science* 68(12):8613–8634.

Yanbc (2024) How are GPU infrastructures managed at OpenAI. Accessed November 30, 2024, https://www.yanbc.info/how-are-gpu-infrastructures-managed-at-openai.

Yang M, Jacob VS, Raghunathan S (2021) Cloud service model's role in provider and user security investment incentives. *Production and Operations Management* 30(2):419–437.

Yenipazarli A (2021) Downstream entry revisited: Economic effects of entry in vertically-related markets. *Omega* 103:102428, ISSN 0305-0483.

Zhang Z, Nan G, Tan Y (2020) Cloud services vs. on-premises software: Competition under security risk and product customization. *Information Systems Research* 31(3):848–864.

Zhao S, Liu B, Wang F, Feng D (2021) Bridging the gap between memory and communication efficiency on distributed deep learning systems. *IEEE Access* 9:57075–57088.