

Cloud or On-Premise? A Strategic View of Large Language Model Deployment

Zhoupeng (Jack) Zhang
University at Buffalo
School of Management
zhoupeng@buffalo.edu

Jiaqi Shi*
University at Buffalo
School of Management
jshi36@buffalo.edu

Shaojie Tang
University at Buffalo
School of Management
shaojiet@buffalo.edu

Abstract

Large language models (LLMs) have advanced rapidly in recent years. We examine a critical decision faced by an LLM provider: whether to provide a local (on-premise) service channel in addition to cloud services. We develop a game-theoretical queueing model to analyze the economic and welfare implications of introducing an on-premise model. Our results show that offering the localization option can reduce the provider's optimal profit due to market cannibalization, yet increase users' overall surplus. Such market outcomes can be reinforced by users' privacy concerns, but may reverse when users differ significantly in their service valuations, as localization enables the provider to extract users' surplus more effectively. When localization is offered through a third party, price discrimination can further increase surplus extraction; however, the double marginalization along the AI supply chain may offset these gains. Finally, in competitive markets, localization may prompt an entrant to lower the quality of their cloud services to limit cannibalization, thereby softening price competition with the incumbent to some extent. Overall, our analysis highlights the strategic trade-offs in LLM deployment and provides guidance on pricing and localization decisions.

Keywords: Large language models, generative AI economics, on-premise deployment, pricing strategies

1. Introduction

The rapid advancement of Large Language Models (LLMs) has ushered in a new era of AI applications across industries, such as user service (Shareef 2024), software development (Lin et al. 2024), and content

generation (Grewal et al. 2025), etc. According to the report by *Research and Markets*, the global LLM market was valued at approximately \$6.33 billion in 2024, reflecting an annual growth rate of 29.85% since 2019. The market is projected to grow to \$25.22 billion by 2029 and reach \$95.45 billion by 2034. These powerful LLM models, like ChatGPT, are initially introduced to the public as cloud-based services, allowing providers to centralize computing resources, control model updates, and gather user data for model training. However, a growing number of users, especially enterprises and governments, demand the ability to run LLMs locally, a practice often called on-premise deployment. This shift is driven by data privacy, cloud latency, and customization concerns (Pan et al. 2020, Hu et al. 2022). As a result, LLM vendors now face a strategic crossroads: Should they continue to offer services solely through the cloud, or should they also enable on-premise deployments that run independently on users' infrastructure? The market is already witnessing divergent strategies; for instance, OpenAI has largely adhered to a cloud-only service, while competitors like Deepseek offer cloud and on-premise solutions. According to Stanford's AI index report, in 2024, 32.8% of AI models were available to the public only through API access. This represents a fundamental trade-off in business model design, raising questions about its implications for firm profitability and user welfare.

Despite the growing literature on the economics of generative AI (Xu et al. 2024, Bergemann et al. 2025) and vendor competition in duopoly marketplaces (Ma and Seidmann 2015, Li et al. 2025), research on the trade-offs between LLM cloud-only and hybrid delivery models (i.e., offering both cloud-based and

* Corresponding author

on-premise solutions) remains limited (Wu et al. 2025). While cloud-based delivery is still the dominant approach for deploying LLMs (Maslej et al. 2025), the rising demand for on-premise deployments introduces a strategic dilemma for providers. On the one hand, localization can expand the reach of the market by appealing to users with strict privacy, security, or latency requirements. On the other hand, it risks cannibalizing cloud-based revenue and limits providers' ability to collect user interaction data for model improvement. LLMs experience positive and negative externalities from increased cloud usage, unlike traditional software. While more users generate valuable data that can enhance model quality, they also increase response latency. Despite the growing number of hybrid deployment models, there is limited theoretical analysis of their impact on the LLM provider's profit.

We contribute to generative AI research by developing a theoretical model to address several interrelated questions. First, we examine the core strategic decision: how does the choice between cloud-only and hybrid deployment models affect LLM provider profitability and user welfare? Second, we investigate the underlying mechanism of our findings: why might deploying an on-premise model increase or decrease the LLM provider's optimal profit? Third, we analyze how key factors, such as users' heterogeneous valuations of LLM services, the intrinsic quality of the models, the market structure, and user privacy concerns, influence the optimal deployment strategy. By modeling user behavior in the LLM market, this research offers a framework for understanding LLM localization's economic and welfare implications.

Our study addresses these research questions by developing a model in a monopolistic LLM market with heterogeneous user usage. The provider first decides whether to offer a cloud-only or hybrid delivery model. In the cloud-only setting, users access a higher-quality cloud-based model by paying subscription and usage-based fees, potentially facing cloud congestion. In the hybrid setting, users also have the option to deploy an on-premise model, which requires a significant upfront hardware investment. The provider sets pricing for subscription and usage, and users self-select into their preferred service. We begin by solving for the provider's optimal profit under each delivery model and comparing the profits in the two models. We then examine how the profit differs in response to changes in user value heterogeneity, the LLM's intrinsic quality, and users' privacy concerns. Our framework directly compares firm profitability and user welfare across deployment strategies and studies the mechanism of the profit difference.

The main analysis yields several interesting findings. First, by accounting for the unique characteristics of LLMs, such as service delays in cloud delivery and the quality disparity between cloud and on-premise models, we find that in a monopoly market, launching an on-premise model reduces the LLM provider's optimal profit due to market cannibalization. At the same time, the on-premise model increases total user surplus. However, when we relax the assumption of homogeneous user valuations, we find that a hybrid strategy offering both cloud and on-premise models can lead to higher profits for the LLM provider than the cloud-only model.

We explore several extensions beyond the baseline model to enrich our analysis further. First, we relax the assumption of homogeneous user valuations by allowing users to differ in how they value identical services. Second, we extend the model to incorporate quality differentiation, including introducing a free, lower-quality service version. Third, we model user privacy concerns associated with cloud-based services. Fourth, we consider a setting where a third-party AI-PC manufacturer provides the localized LLM rather than the original LLM provider. Finally, we briefly discuss the decision to localize LLM in a competitive market environment. These extensions offer a more comprehensive understanding of LLM localization given diverse market conditions.

Our research contributes to the emerging literature on the economics of generative AI by developing the first formal framework to analyze the trade-off between cloud-only and hybrid LLM deployment strategies. We systematically examine how on-premise deployment affects provider profitability and user surplus by explicitly modeling cloud congestion and related externalities. Our core analysis highlights how on-premise offerings influence both the cloud congestion effect and the market cannibalization effect. Furthermore, our extensions show how user valuation heterogeneity, market structure, and users' privacy concerns shape this trade-off. Overall, our findings provide strategic guidance to LLM providers in making deployment and pricing decisions while offering insights into LLM localization's broader welfare implications.

2. Literature Review

Our work is in connection with the nascent literature on the generative AI economic analysis (Xu et al. 2024, Laufer et al. 2024, Mahmood 2024, Bergemann et al. 2025, Yang and Zhu 2025, Wu et al. 2025, Taitler and Ben-Porat 2025, Yang 2025, Xu et al.

2025). For instance, [Mahmood \(2024\)](#) studies pricing competition between generative AI firms as a sequential game. The author finds that latecomers can often undercut first-movers on price unless the first-mover has strong task-specific performance advantages. [Xu et al. \(2024\)](#) models the foundation model value chain and identifies the “openness trap” phenomenon, where moderate openness can reduce overall welfare. The study also shows that vertical integration and free trials may backfire under specific conditions. [Bergemann et al. \(2025\)](#) presents an economic model for optimal pricing and design of LLMs, accounting for token costs, fine-tuning, and user heterogeneity. The analysis reveals that two-part tariffs can efficiently segment users by usage intensity and customization preferences, aligning well with current industry practices. [Yang and Zhu \(2025\)](#) proposes a pricing framework for agentic AI services that balances cost, quality, and user incentives under asymmetric information. This stream of literature primarily focuses on pricing issues for cloud-based AI products without accounting for the deployment of free on-premises models or their economic implications. Closest to our setting, [Wu et al. \(2025\)](#) analyze users’ adoption incentives for closed-source APIs versus open-source models, emphasizing the role of community engagement in sustaining innovation. Distinct from this literature, our study focuses on the provider’s dilemma of whether to introduce an on-premises option, and its consequences for profitability and welfare in monopoly markets.

Beyond AI-specific work, our paper also builds on research examining strategic and operational trade-offs between Software-as-a-Service (SaaS) and on-premise software. Unlike traditional software, SaaS involves a subscription model in which vendors host and maintain software on their hardware ([Choudhary 2007](#)). Operational strategies studied in SaaS research include vendors’ pricing, usage regulation, cost optimization, capacity expansion, investments in software quality and security, and so forth ([August et al. 2014](#), [Zhang et al. 2020](#), [Li and Kumar 2022](#), [Chen et al. 2023](#), [Qu et al. 2024](#), [Li et al. 2025](#)). For example, [Choudhary \(2007\)](#) explores how SaaS licensing incentivizes greater software quality investment than perpetual licensing, resulting in higher profits and improved social welfare. Regarding security, [Zhang et al. \(2020\)](#) provides the counterintuitive finding that SaaS yields lower expected user losses than on-premise software in scenarios of significant potential losses from security incidents. However, SaaS vendors exhibit limited motivation to enhance security measures in environments with low potential security losses, whereas on-premises vendors tend to increase their security investments.

Conversely, studies such as [August et al. \(2014\)](#) indicate that users have transparency and compliance concerns regarding SaaS, leading them to prefer on-premises over SaaS alternatives. The issue of SaaS customizability is addressed in [Li et al. \(2025\)](#), which reveals that increased customizability does not necessarily benefit SaaS vendors in a duopoly market. This stream of literature typically treats SaaS and on-premises software as substitutes, comparing which model dominates under different conditions. We extend this perspective by analyzing hybrid deployment in the context of generative AI, where cloud delay, LLM quality difference, and user fine-tuning create distinctive trade-offs for providers.

Finally, our study connects with the literature on revenue management and pricing of information goods. Foundational work emphasizes the cost structure of digital goods ([Varian 1995](#), [Bakos and Brynjolfsson 1999](#), [Huang and Sundararajan 2011](#)) and develops strategies such as subscriptions, pay-per-use, bundling, and dynamic pricing ([Gallego and van Ryzin 1994](#), [Fishburn and Odlyzko 1999](#), [Wu et al. 2008](#), [Balasubramanian et al. 2015](#), [Chen and Chen 2015](#)). Within SaaS markets, in addition to the research of pricing strategies ([Li and Kumar 2022](#)), the research further explores the price competition between SaaS and traditional vendors ([Ma and Seidmann 2015](#), [Li et al. 2025](#)). [Ma and Seidmann \(2015\)](#) studies competition between SaaS and traditional software providers, recommending that SaaS vendors reduce lack-of-fit costs and lower per-transaction prices to exploit economies of scale. In contrast, traditional vendors are advised to avoid competing on price and instead focus on feature enhancements. In a duopoly setting, [Li et al. \(2025\)](#) finds that SaaS vendors typically prefer pay-per-use pricing over subscription models, as it broadens the user base and mitigates competitive pressure from on-premises providers. Building on these insights, we develop a theoretical framework tailored to LLM providers, showing how the option to offer on-premises services reshapes optimal pricing, surplus distribution, and monopoly profits.

3. Model

Consider the operations of a single LLM provider that provides the AI services over infinitely many time periods, $t = 1, 2, \dots$ Both the provider and LLM users discount future outcomes at a rate of $\delta \in (0, 1)$. Users are heterogeneous and mainly differentiated in the volume of their demand for LLM services: Some frequently solicit AI-generated content, while others do not. To capture this, we assume that in each time

period t , an infinitesimal user has a usage rate τ , which measures the number of queries (or prompts) he or she sends to the LLM and is drawn independently from $U[0, 1]$. We normalize the population of users N to 1.

There are two ways for a user to access the LLM services, *cloud* and *local*. Users make their channel decisions prior to the beginning of the time horizon (i.e., at $t = 0$). The long-run utility of a user with an LLM usage rate of τ is $u_c = (\tau(v - cW(\lambda, \mu) - p) - F)/(1 - \delta)$, where v is the average valuation of each service, c is marginal disutility of service delay, $W(\lambda, \mu)$ is the average service delay on cloud, μ is the LLM's service capacity, λ is the (expected) total traffic on cloud and, finally, p and F are the per-usage and the flat fee for the cloud service in each time period, respectively. One can think of p as the price for API, which is typically charged on top of the subscription fee F .¹ Note that λ is endogenous to users' channel choices: the more users choose cloud, the larger the λ , and vice versa. A formal definition of λ will be provided below. For the ease of exposition, we formulate the average service delay as follows: $W(\lambda, \mu) = (\lambda/\mu)/(\mu - \lambda)$ for $\lambda < \mu$ and $W(\lambda, \mu) = \infty$ otherwise, which, as one may have noticed, is the average wait time formula for an $M(\lambda)/M(\mu)/1$ queue. Despite its stylized queuing nature, such a formulation is consistent with the IS literature on service delay in, e.g., the Internet service industry (see, e.g., Cheng et al. 2011). Our key insights continue to hold with alternative formulations, e.g., $W(\lambda, \mu) = (\lambda/\mu)^{\gamma(s)}/(\mu - \lambda)$, the approximated service delay in a multi-server $M(\lambda)/M(\mu)/s$ queue. Also note that the formulation of $W(\lambda, \mu)$ implies that users are sensitive only to the time it takes for the LLM to serve the other users, but not to that for themselves (otherwise, $W(\lambda, \mu)$ shall be $(\lambda/\mu)/(\mu - \lambda) + 1/\mu = 1/(\mu - \lambda)$, where $1/\mu$ is the average service time for a focal user). Our main results can nevertheless be generalized.

If a user instead chooses to localize the AI service, his or her utility long-run becomes $u_l = \tau \cdot (\alpha(K)v - c/\mu_l(\kappa))/(1 - \delta) - K - \kappa$, where K is the initial investment of localization and $\alpha(K)$ is the capability coefficient of LLM localization. Here, K includes the setup expense, fine-tuning effort, etc., and is controlled by users. In Section 5.4, we consider an alternative market structure where the localized service is also provided by a third-party company (e.g., a manufacturer of AI PC); there, K is the equipment fee r charged by the company. We assume $\alpha'(K) > 0$ and $\alpha''(K) < 0$. That $\alpha'(K) > 0$ implies that the more users invest in localization, the more capable the local AI will be, and that $\alpha''(K) < 0$ indicates that such a positive return

is nevertheless marginally diminishing. To facilitate the analysis, we assume $\alpha(K) = \sqrt{K}$. As such, for a user with usage rate τ , should he or she choose localization, the optimal amount of investment $K^* = (\tau v/(1 - \delta))^2/4$, and the maximal local utility is $u_l^* = (\tau v/(1 - \delta))^2/4$.

We are now ready to characterize the user equilibrium given service prices p and F . Clearly, for users whose cloud utility u_c is higher than their maximal local utility u_l^* , they will choose the cloud channel, and vice versa. Define $\Delta = u_c - u_l^* = (\tau(v - cW(\lambda, \mu) - p) - F)/(1 - \delta) - (\tau v/(1 - \delta))^2/4$, which is a quadratic function in the usage rate τ . Define $\underline{\tau}, \bar{\tau}$ as the root of Δ . Suppose both $\underline{\tau}$ and $\bar{\tau}$ exist and that $0 \leq \underline{\tau} < \bar{\tau} \leq 1$. Then by the definition of Δ , users whose usage rate $\tau \in [\underline{\tau}, \bar{\tau}]$ will have a $\Delta \geq 0$ and thus choose cloud, and the rest of them will have a $\Delta < 0$ and pick local. Therefore, the total number of cloud users is $\bar{\tau} - \underline{\tau}$, and the cloud traffic $\lambda = \int_{\underline{\tau}}^{\bar{\tau}} x dx = (\bar{\tau}^2 - \underline{\tau}^2)/2$.

In light of the significant energy consumption during an LLM's operations, we assume that processing each user request on average incurs a cost of $\omega > 0$. The LLM provider's long-run profit can thus be characterized as $\Pi = ((p - \omega) \cdot (\bar{\tau}^2 - \underline{\tau}^2)/2 + F \cdot (\bar{\tau} - \underline{\tau}))/ (1 - \delta)$, and its problem is to find out the optimal pair of p and F that maximizes Π . We also define users' total surplus $US \equiv \int_0^{\underline{\tau}} (xv/(1 - \delta))^2/4 dx + \int_{\underline{\tau}}^{\bar{\tau}} (x(v - cW(\lambda, \mu) - p) - F)/(1 - \delta) dx + \int_{\bar{\tau}}^1 (xv/(1 - \delta))^2/4 dx$.

Some remarks about our theoretical framework assumptions are in order. First, we have assumed that the quality of LLM output is constant and homogeneous across users, which is admittedly simplistic. In reality, quality differentiation is a common practice in the generative AI industry. For example, by the time we are writing this paper, ChatGPT has a model lineup of GPT 4.1-mini, GPT 4.1, GPT 4o, and GPT 4.5, among others, with the first model being accessible by Free users and the other three models being used mostly by Plus and Pro users. That said, our theoretical framework can be extended in this way, and our main findings and insights are robust. Second, we have assumed away any horizontal mismatches between LLM's outputs and users' idiosyncratic preferences, which is nonetheless a critical issue in practice. Indeed, an important motivation for people to localize an LLM is that they can better fine-tune the model and tailor the AI-generated content to their own needs. At a high level, incorporating this aspect of reality into our theoretical framework makes the option of taking the LLM local even more attractive to users, thus aggravating the negative impacts of localization on the provider's cloud

¹See, e.g., <http://openai.com/api/pricing/>.

businesses.

Next, we have abstracted away the provider's decisions to upgrade the LLM's service quality (i.e., increase v) and expand the cloud service capacity (i.e., increase μ) over time. The generative AI industry is, however, marked by incessant waves of model training/upgrades and infrastructure enhancement. As our discussion below would imply, by allowing users to localize the LLM, the LLM provider is essentially creating itself a competitor, and this will weaken or even reverse the positive effects of LLM upgrades and cloud capacity expansions, thus souring the provider's financial interests in these moves. Finally, one may notice that users in our theoretical framework settle their service access channels once and for all. That is, they will not switch back and forth between cloud and local from one period to another or mix their usages across channels in any single time period. Rather than being an assumption per se, this follows our theoretical framework setup naturally: One can verify mathematically that given users' channel choices at the beginning, there will be no incentive for any of them to switch later on whatsoever.

3.1. Benchmark: No Localization

To benchmark our analysis of a market with both cloud and local services, here we set up a base scenario where no localization is allowed. That is, users who wish to access LLM services must do so via the cloud. Note that in each period, by subscribing to cloud services, a user collects a utility of $\tau(v - c(\lambda_0/\mu)/(\mu - \lambda_0) - p) - F$, where λ_0 denotes the equilibrium cloud traffic in this no-localization benchmark. As such, only users with a usage rate of $\tau \geq \tau_0 \equiv F/((v - c(\lambda_0/\mu)/(\mu - \lambda_0) - p))$ will subscribe. Suppose $\tau_0 \in [0, 1]$. Then the number of subscribers is $1 - \tau_0$, and the cloud traffic is $\lambda_0 = \int_{\tau_0}^1 x dx = (1 - \tau_0^2)/2$. The LLM provider's long-run profit can thus be characterized as $\Pi_0 = ((p - \omega) \cdot (1 - \tau_0^2)/2 + F \cdot (1 - \tau_0))/(1 - \delta)$, and its problem is to find out the optimal pair of p and F that maximizes Π_0 . Users' total surplus in this case is defined as $US_0 \equiv \int_{\tau_0}^1 (x(v - cW(\lambda, \mu) - p) - F)/(1 - \delta) dx$.

4. Analysis

To understand how localization affects the LLM provider's operations, it is useful to first pin down the market equilibrium for the no-localization benchmark. The lemma below characterizes the LLM provider's optimization problem in that benchmark.

Lemma 1 *When users are not allowed to localize the LLM, the provider's profit maximization problem over p and F is equivalent to an optimal control problem over the amount of cloud traffic λ_0 in each period: Given the desired amount of cloud traffic $\lambda_0 = (1 - \tau_0^2)/2$, the provider will set the per-usage fee $p = v - c(\lambda_0/\mu)/(\mu - \lambda_0)$ and subscription fee $F_0 = 0$. The optimal control problem is as follows,*

$$\max_{\lambda_0 \in [0, \frac{1}{2}]} \Pi_0 = \lambda_0(v - \frac{c(\lambda_0/\mu)}{\mu - \lambda_0} - \omega)/(1 - \delta). \quad (1)$$

Lemma 1 shows that the operations of an LLM provider in the short run (i.e., in each time period) are mainly about maintaining a financially optimal scale of cloud traffic. This requires the provider to balance two competing forces: On the one hand, more traffic creates extra sources of revenue, yet on the other hand, it also aggravates the congestion on the cloud, extends service delays, and ultimately reduces users' willingness to pay. Note that, as per the definition of cloud traffic $\lambda_0 = (1 - \tau_0^2)/2$, we have a one-to-one mapping between λ_0 and the usage cutoff τ_0 . As such, the optimization problem (1) can also be thought of as an optimal control problem over τ_0 .

By solving the problem (1), we can back out the no-localization equilibrium as follows.

Proposition 1 *When users are not allowed to localize the LLM, the provider's optimal profit in the long-run $\Pi_0^* = \lambda_0^*(v - c(\lambda_0^*/\mu)/(\mu - \lambda_0^*) - \omega)/(1 - \delta)$, where the optimal cloud traffic in each period $\lambda_0^* = ((v - \omega)\mu + c - \sqrt{c((v - \omega)\mu + c)})/(v - \omega + c/\mu)$ if $\mu \leq \bar{\mu}_0$ for some $\bar{\mu}_0$ and $\lambda_0^* = 1/2$ otherwise. In each period, users with a usage rate $\tau \geq \tau_0^* = \sqrt{1/2 - \lambda_0^*}$ will subscribe, and users' total surplus $US_0^* = 0$.*

We would like to highlight that because all users receive a utility of $v - c(\lambda_0^*/\mu)/(\mu - \lambda_0^*)$ from the LLM each time, the provider will extract all their surplus by charging a per-usage fee $p_0^* = v - c(\lambda_0^*/\mu)/(\mu - \lambda_0^*)$ as per Lemma 1. Therefore, users' total surplus will be zero in the equilibrium.

We are now ready to dive into the scenario where users are allowed to localize the LLM. Recall that given the provider's prices p and F , in general, there exist two thresholds $\underline{\tau}$ and $\bar{\tau}$ on users' usage rate τ such that $\underline{\tau} \leq \bar{\tau}$ and that users will choose cloud if and only if $\tau \in [\underline{\tau}, \bar{\tau}]$, and vice versa.

Lemma 2 *When users are allowed to localize the LLM, the firm's profit maximization problem over p and F is equivalent to an optimal control problem over usage thresholds $\underline{\tau}$ and $\bar{\tau}$ in each period: Given desired thresholds $\underline{\tau}$ and $\bar{\tau}$ such that $0 \leq \underline{\tau} \leq \bar{\tau} \leq 1$, the*

provider will set the per-usage fee $p = v - c(\lambda/\mu)/(\mu - \lambda) - v^2(\bar{\tau} + \underline{\tau})/(4(1 - \delta))$ and subscription fee $F = v^2\bar{\tau}\underline{\tau}/(4(1 - \delta))$, where the cloud traffic $\lambda = (\bar{\tau}^2 - \underline{\tau}^2)/2$. The optimal control problem is,

$$\max_{0 \leq \bar{\tau} \leq \underline{\tau} \leq 1} \Pi = \left(\frac{\bar{\tau}^2 - \underline{\tau}^2}{2} \cdot \left(v - \frac{c(\bar{\tau}^2 - \underline{\tau}^2)/(2\mu)}{\mu - (\bar{\tau}^2 - \underline{\tau}^2)/2} - \omega - \frac{v^2(\bar{\tau}^2 + \underline{\tau}^2)}{4(\bar{\tau} + \underline{\tau})(1 - \delta)} \right) \right) / (1 - \delta). \quad (2)$$

Similar to Lemma 1, Lemma 2 says that the provider's operations in each period are essentially controlling how many people would subscribe and go "cloud." However, the problem (2) in Lemma 2 also reveals the twist localization has added to the picture: It boosts users' outside options from zero (i.e., not using the LLM at all) to something positive (i.e., using the LLM locally at some initial setup costs) and therefore lowers everybody's willingness to pay for the cloud services. Indeed, comparing the provider's profit function in (2) and that in (1), one can see that the average cloud service revenue is lower by $(v^2(\bar{\tau}^2 + \underline{\tau}^2))/(4(\bar{\tau} + \underline{\tau})(1 - \delta))$. In other words, by allowing users to localize the LLM, the provider creates itself a competitor. We dub this the *cannibalization effect*.

Albeit unavailable in closed form, we can analyze the market equilibrium based on Lemma 2 and compare it with that in the no-localization baseline. The proposition below presents our findings.

Proposition 2 *Localization decreases the LLM provider's optimal profit (i.e., $\Pi_0^* \geq \Pi^*$) but increases users' total surplus (i.e., $US_0^* \leq US^*$).*

Straightforward as it may seem, we shall elaborate on Proposition 2, especially its part on the provider's profit. Recall the cannibalization effect of localization we introduced earlier. Suppose the provider starts in the no-localization benchmark, but at some point in time, it turns on the localization option. Because of the cannibalization effect, some users will switch from cloud to local. Though immediately hitting the provider's cloud traffic, it nevertheless alleviates the congestion on the cloud, thus improving the service experience for users who stay behind and increasing their willingness to pay. Then why is such a positive *congestion effect* necessarily dominated by the cannibalization effect (and thus the provider ends up with a lower profit)? The intuition is that while the congestion effect works only on users who continue with the cloud, the cannibalization effect tumbles the revenue the provider collects from everyone. In particular, for users who stay on the cloud, the cannibalization effect cancels out a significant portion of their increased

willingness to pay thanks to the congestion effect, thus making its overall positive impact minimal, if at all.

It is also worthwhile to note that such a profit gap caused by localization would enlarge as the LLM's service quality (or capability) v increases. See Figure 1 for an illustration. Indeed, while a higher v always increases the LLM provider's optimal profit in the no-localization benchmark, it initially increases but then decreases the profit when users are given the localization option. This implies that localization can also slow down the LLM provider's model upgrades over time, as the financial return on such investments falters due to localization's cannibalization.

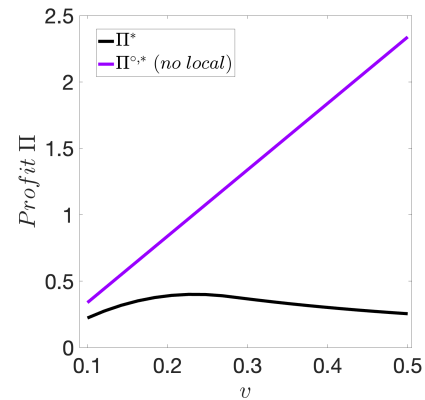


Figure 1: LLM's optimal profit as service valuation changes ($\mu = 15, c = 1, \omega = 0.03, \delta = 0.9$)

On the other hand, despite the cannibalization effect, the positive impact of an expanded cloud capacity remains relatively unchanged when localization is introduced. As Figure 2 has illustrated, the marginal increases in the provider's profit in response to a higher μ are similar with and without localization. The intuition is that, for cloud users, with or without localization, an additional unit of cloud capacity $\Delta\mu$ always generates relatively the same amount of increase in their utility u_c . Such utility gains eventually translate into extra revenue of similar size for the provider.

5. Extensions

5.1. Valuation Heterogeneity

We have so far assumed that users perceive the LLM service quality identically, though in reality, they may differ in their valuation of the same service. To capture such heterogeneity, suppose an infinitesimal user perceives the service quality to be θv , where v represents the average perceived quality and θ i.i.d. $\sim U[1/2, 3/2]$ is the idiosyncratic valuation parameter.

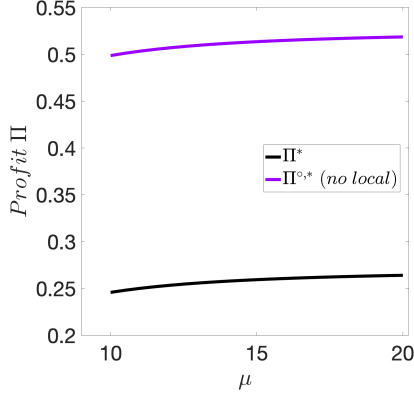


Figure 2: LLM's optimal profit as service capacity changes ($v = 0.15, c = 1, \omega = 0.05, \delta = 0.9$)

To single out the impact of valuation heterogeneity, suppose users share the same usage rate $\tau = 1$. We can show that the LLM provider's optimization problem without localization is:

$$\max_{v_0 \in [\frac{v}{2}, \frac{3v}{2}]} \frac{(\frac{3v}{2} - v_0) \cdot (v_0 - \frac{c(\frac{3v}{2} - v_0)/\mu}{\mu - (\frac{3v}{2} - v_0)} - \omega)}{1 - \delta}. \quad (3)$$

And that with localization becomes:

$$\max_{\frac{v}{2} \leq v < \bar{v} \leq \frac{3v}{2}} \frac{(\bar{v} - v)}{1 - \delta} \cdot \left(\frac{v + \bar{v}}{2} - \frac{c(\bar{v} - v)/\mu}{\mu - (\bar{v} - v)} \right) - \omega - \frac{(v^2 + \bar{v}^2)}{8(1 - \delta)}. \quad (4)$$

Compare (3) with (4). While the cannibalization effect, embodied by the term $-\tau(v^2 + \bar{v}^2)/(8(1 - \delta))$ in (4), still takes a toll on the LLM provider's long-run profit, a new positive effect now enters the scene: Localization (surprisingly) enables the provider to extract relatively more surplus from cloud users. Indeed, without localization, the amount of user surplus the provider extracts for each service is just v_0 (see (3)), the lowest valuation among all cloud users. In contrast, with localization, the provider extracts $(v + \bar{v})/2$ (see (4)), the average between the highest and the lowest valuations among all cloud users. As such, localization may now benefit the provider but hurt users.

5.2. Differentiated Service Quality

As mentioned in Section 3, our baseline assumes that the LLM provider offers only a single quality tier for LLM services; however, in reality, quality differentiation is a common practice. LLM providers typically offer a basic model for free and some premium

models for a fee. To incorporate this into our framework, suppose there are two quality tiers on the cloud, one base tier with quality v_L and one premium tier with quality $v_H > v_L$. Users can use the base model for free, but have to pay the subscription fee F together with the per-usage fee p to access the premium model.

At a high level, localization serves as an additional option for users. If localization is less attractive than continuing to use the free cloud model, then the users' problem remains choosing between the free cloud model and the premium cloud model, and the market equilibrium will be similar to that without localization. Otherwise, the users' problem becomes choosing between localization and the premium cloud model. As such, the cannibalization effect is again at play, and the LLM provider will be hurt while users will benefit from localization.

5.3. Users' Privacy Concerns and Data Spillover

One of the users' primary concerns with cloud-based LLMs, and AI more broadly, is the risk of data leakage and spillover (Pan et al. 2020, Tucker et al. 2018). In a cloud service setting, all user inputs must be transmitted to the provider's data center, which introduces the risk of data breaches in a cyberattack. Moreover, proprietary data from one company may be memorized by the model and potentially surface in responses to queries from competitors. To formally capture this concern, we incorporate a disutility term into the user's utility function for cloud-based services, representing the perceived risk of privacy loss.

Introducing a negative term in the cloud utility function to account for privacy concerns widens the provider's potential profit gap between the hybrid and cloud-only models, regardless of the distribution of privacy sensitivity among users. This occurs because the added disutility amplifies the cannibalization effect: it pushes an additional segment of privacy-conscious users toward the on-premise model, thereby reducing demand for the cloud service.

5.4. Third-Party Provision of Localization

Our baseline presumes that users localize the LLM on their own (e.g., downloading the LLM to their current PC). Now, consider the case where localized services are provided by an AI-PC manufacturer. For simplicity, suppose purchasing an AI-PC is the only way for localization. Future research could study hybrid scenarios with AI-PCs and DIY localization.

Suppose the manufacturer charges r for each AI-PC and pays back κ to the LLM provider. In addition,

the per-unit manufacturing cost is m . To comprehend how the market will unravel in this setting, first consider the vertically integrated scenario where the LLM provider has full control over the manufacturer's price r . Then, localization serves as a price discrimination device, enabling the provider to sell to users with different preferences for cloud versus local services at different prices. As such, localization shall enhance the provider's profitability.

In the disintegrated scenario where the manufacturer independently sets his price, the classic double marginalization effect would distort the AI-PC price r upward so that the total number of AI-PCs sold will be fewer than what the LLM provider would prefer. Hence, the ultimate impact of LLM on the LLM provider hinges on the relative magnitude of this negative double marginalization effect and the positive price discrimination effect.

5.5. Localization in a Competitive Market

We now place localization in the context of a competitive generative AI market. Suppose an incumbent (e.g., ChatGPT) has been offering closed-source and cloud-only services to users. Now, an entrant (e.g., DeepSeek) is contemplating market entry and, importantly, whether to make its own LLM open-source and give users the option of localization. If the entrant seeks market penetration and rapid expansion in the short term, then enabling localization can be an effective strategy, as the benefits of zero service delay (and of local fine-tuning, etc.) help lure customers away from the incumbent's closed-source LLM on the cloud.

If the entrant instead targets profitability, then the impact of localization is more nuanced. On one hand, the incumbent is poised to lower her price to avoid users flocking to the entrant's local channel, which will add downward pressure on the entrant's price for his own cloud services. On the other hand, localization also changes the entrant's optimal quality decision. As shown in Figure 1, when localization is available, beyond a certain point, raising quality v intensifies the cannibalization of an LLM provider's cloud business and reduces overall profit. Anticipating this, the entrant may strategically choose a lower cloud service quality to limit cannibalization. This leads to greater vertical differentiation between the incumbent and the entrant, and as a result, price competition may become less intense.

6. Conclusions

In this paper, we develop a theoretical framework to analyze the economic and strategic trade-offs faced by

LLM providers when choosing between cloud-only and hybrid (cloud and on-premise) deployment strategies. We model the provider's optimal profit under each strategy by incorporating factors such as cloud congestion and model quality differences. Our analysis reveals that while enabling localization may reduce the provider's optimal profit due to the cannibalization of its cloud service, it enhances overall user welfare. Significantly, this adverse profit effect can be reversed when users have heterogeneous valuations of the service.

While the fundamental trade-off revealed in this paper might resemble those in traditional cloud or software services, LLMs introduce unique challenges that our current framework does not fully capture. In conventional cloud systems, latency and resource consumption scale predictably with the number of users, whereas LLMs exhibit highly uneven computational demands: a small set of complex prompts can generate far greater congestion than thousands of simple queries, and users issuing such complex requests often tolerate longer delays. This heterogeneity in both task complexity and user patience breaks the classic assumptions of homogeneous demand in cloud models. Extending the framework to incorporate these dynamics, along with factors such as fine-tuning costs, privacy concerns, and variation in user expertise, would provide a richer understanding of the economic and welfare implications of hybrid LLM deployment.

7. References

- August T, Niculescu MF, Shin H (2014) Cloud implications on software network structure and security risks. *Information Systems Research* 25(3):489–510.
- Bakos Y, Brynjolfsson E (1999) Bundling information goods: Pricing, profits, and efficiency. *Management Science* 45(12):1613–1630.
- Balasubramanian S, Bhattacharya S, Krishnan VV (2015) Pricing information goods: A strategic analysis of the selling and pay-per-use mechanisms. *Marketing Science* 34(2):218–234.
- Bergemann D, Bonatti A, Smolin A (2025) The economics of large language models: Token allocation, fine-tuning, and optimal pricing. *Proceedings of the 26th ACM Conference on Economics and Computation*, 786 (New York, NY, USA: Association for Computing Machinery).
- Chen M, Chen ZL (2015) Recent developments in dynamic pricing research: multiple products, competition, and limited demand information. *Production and Operations Management* 24(5):704–731.
- Chen S, Moinszadeh K, Song JS, Zhong Y (2023) Cloud computing value chains: Research from the operations management perspective. *Manufacturing & Service Operations Management* 25(4):1338–1356.
- Cheng HK, Bandyopadhyay S, Guo H (2011) The debate on net neutrality: A policy perspective. *Information Systems Research* 22(1):60–82.

- Choudhary V (2007) Comparison of software quality under perpetual licensing and software as a service. *Journal of management information systems* 24(2):141–165.
- Fishburn PC, Odlyzko AM (1999) Competitive pricing of information goods: Subscription pricing versus pay-per-use. *Economic Theory* 13:447–470.
- Gallego G, van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* 40(8):999–1020.
- Grewal D, Saturnino CB, Davenport T, Guha A (2025) How generative AI is shaping the future of marketing. *Journal of the Academy of Marketing Science* 53:702–722.
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W, et al. (2022) LoRA: Low-rank adaptation of large language models. *ICLR* 1(2):3.
- Huang KW, Sundararajan A (2011) Pricing digital goods: Discontinuous costs and shared infrastructure. *Information Systems Research* 22(4):721–738.
- Laufer B, Kleinberg J, Heidari H (2024) Fine-tuning games: Bargaining and adaptation for general-purpose models. *Proceedings of the ACM Web Conference 2024*, 66–76 (New York, NY, USA: Association for Computing Machinery).
- Li B, Kumar S (2022) Managing software-as-a-service: Pricing and operations. *Production and Operations Management* 31(6):2588–2608.
- Li N, Jia K, Juan F (2025) SaaS or on-premises? compete through customizability, price, and hybrid offerings. *Production and Operations Management* 34(9):2742–2757.
- Lin F, Kim DJ, et al. (2024) When LLM-based code generation meets the software development process. *arXiv e-prints* arXiv:2403.2403.
- Ma D, Seidmann A (2015) Analyzing software as a service with per-transaction charges. *Information Systems Research* 26(2):360–378.
- Mahmood R (2024) Pricing and competition for generative AI. *Advances in Neural Information Processing Systems* 37:75727–75748.
- Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, Capstick E, Reuel A, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Niebles JC, Shoham Y, Wald R, Walsh T, Hamrah A, Santarlasci L, Lotufo JB, Rome A, Shi A, Oak S (2025) The AI index 2025 annual report. Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.
- Pan X, Zhang M, Ji S, Yang M (2020) Privacy risks of general-purpose language models. *2020 IEEE Symposium on Security and Privacy (SP)*, 1314–1331.
- Qu Z, Dawande M, Janakiraman G (2024) Cloud cost optimization: Model, bounds, and asymptotics. *Operations Research* 72(1):132–150.
- Shareef F (2024) Enhancing conversational AI with LLMs for customer support automation. *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, 239–244.
- Taitler B, Ben-Porat O (2025) Braess’s paradox of generative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 14139–14147.
- Tucker C, Agrawal A, Gans J, Goldfarb A (2018) Privacy, algorithms, and artificial intelligence. *The Economics of Artificial Intelligence: An Agenda*, 423–437 (Chicago: University of Chicago Press).
- Varian HR (1995) Pricing information goods. *Proceedings of Scholarship in the New Information Environment Symposium*, 1–7 (Harvard Law School).
- Wu Sy, Hitt LM, Chen Py, Anandalingam G (2008) Customized bundle pricing for information goods: A nonlinear mixed-integer programming approach. *Management Science* 54(3):608–622.
- Wu Y, Duan H, Li X, Hu X (2025) Navigating the deployment dilemma and innovation paradox: Open-source versus closed-source models. *Proceedings of the ACM on Web Conference 2025*, 1488–1501.
- Xu F, Hou J, Chen W, Xie K (2025) Generative AI and organizational structure in the knowledge economy. Working paper, University of Connecticut, Fudan University.
- Xu F, Wang X, Chen W, Xie K (2024) The economics of AI foundation models: Openness, competition, and governance. Working paper, University of Connecticut, Hong Kong Polytechnic University.
- Yang G (2025) Agentic AI: Service operations with augmentation and automation AI. Working Paper.
- Yang YT, Zhu Q (2025) PACT: A contract-theoretic framework for pricing agentic AI services powered by large language models. arXiv preprint arXiv:2505.21286.
- 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.

Proof of Proposition 2

Because the LLM provider’s long-run profit is equal to the per-period profit weighted by $1/(1-\delta)$, it suffices to analyze the provider’s optimization problem in each period. As such, we drop the term $1/(1-\delta)$ henceforth.

No-Localization Benchmark. Denote by $\tau_0 \in [0, 1]$ the equilibrium usage cutoff given some per-usage price p and subscription fee F . The corresponding total cloud traffic $\lambda_0 = (1 - \tau_0^2)/2$, and those *marginal* users who are indifferent between subscribing and not subscribing have a utility of $u = \tau_0(v - c(\lambda_0/\mu))/(\mu - \lambda_0) - p) - F = 0$. The total number of subscribers is $N_0 = 1 - \tau_0$.

Conversely, should the LLM wish to induce a cloud traffic of λ_0 units, it must ensure that the per-usage price p and subscription fee F satisfy $\tau_0(v - c(\lambda_0/\mu))/(\mu - \lambda_0) - p) - F = 0$, or equivalently, $F = \tau_0(v - c(\lambda_0/\mu))/(\mu - \lambda_0) - p$. As such, the LLM’s total profit will be $\Pi_0 = (p - \omega) \cdot \lambda_0 + F \cdot N_0 = (1 - \tau_0)(\tau_0(v - c(\lambda_0/\mu))/(\mu - \lambda_0)) - \omega \cdot (1 + \tau_0)/2 + p \cdot (1 - \tau_0)^2/2$. One can see that the profit Π is increasing in the per-usage charge p . As such, to maximize the profit, the LLM provider must set p as high as possible. One can easily verify that the highest possible value of p is $v - c(\lambda_0/\mu)/(\mu - \lambda_0)$ as otherwise subscribers’ utility $\tau(v - c(\lambda_0/\mu))/(\mu - \lambda_0) - p) - F$ will become negative for any non-negative subscription fee F . As such, the

optimal $p^* = v - c(\lambda_0/\mu)/(\mu - \lambda_0)$. Plug p^* into Π_0 and after some algebraic manipulation, we have $\Pi_0 = (1 - \tau_0) \cdot ((v - c(\lambda_0/\mu)/(\mu - \lambda_0))(1 + \tau_0)/2 - \omega(1 + \tau_0)/2) = ((1 - \tau_0^2)/2) \cdot (v - c(\lambda_0/\mu)/(\mu - \lambda_0) - \omega)$. As such, the profit maximization problem is essentially an optimal control problem over the usage cutoff τ_0 , which can be characterized as

$$\max_{\tau_0 \in [0,1]} \frac{1 - \tau_0^2}{2} \cdot \left(v - \frac{c(1 - \tau_0^2)/(2\mu)}{\mu - \frac{1 - \tau_0^2}{2}} - \omega \right). \quad (5)$$

With Localization. We now turn to the case where the LLM allows users to localize. We know that there are two cutoffs $\underline{\tau}$ and $\bar{\tau}$ such that $0 < \underline{\tau} \leq \bar{\tau} \leq 1$ and that users with $\tau \in [\underline{\tau}, \bar{\tau}]$ will choose cloud or switch to local otherwise. As such, the cloud traffic will be $\lambda = (\bar{\tau}^2 - \underline{\tau}^2)/2$, and the number of subscribers is $N = \bar{\tau} - \underline{\tau}$.

Given that users with $\underline{\tau}$ or $\bar{\tau}$ must be indifferent between cloud and local, we have

$$\frac{(v\underline{\tau})^2}{4(1 - \delta)} - \underline{\tau}(v - c\frac{\lambda/\mu}{\mu - \lambda} - p) + F = 0, \quad (6)$$

$$\frac{(v\bar{\tau})^2}{4(1 - \delta)} - \bar{\tau}(v - c\frac{\lambda/\mu}{\mu - \lambda} - p) + F = 0. \quad (7)$$

Taking the difference between (6) and (7) and we get $p = v - c(\lambda/\mu)/(\mu - \lambda) - v^2(\bar{\tau} + \underline{\tau})/(4(1 - \delta))$. Plug this p into either (6) or (7), we get $F = v^2\bar{\tau}\underline{\tau}/(4(1 - \delta))$. We can thus transform the LLM's total profit $\Pi = (p - \omega) \cdot \lambda + F \cdot N = ((\bar{\tau}^2 - \underline{\tau}^2)/2) \cdot (v - c(\bar{\tau}^2 - \underline{\tau}^2)/(2\mu)/(\mu - (\bar{\tau}^2 - \underline{\tau}^2)/2) - \omega - v^2(\bar{\tau}^2 + \underline{\tau}^2)/(4(\bar{\tau} + \underline{\tau})(1 - \delta)))$. One can see that the profit maximization problem is equivalent to an optimal control problem over $\bar{\tau}$ and $\underline{\tau}$.

Now, notice that to maximize the profit Π over $\bar{\tau}$ and $\underline{\tau}$, we have

$$\begin{aligned} & \max_{0 \leq \underline{\tau} \leq \bar{\tau} \leq 1} \frac{\bar{\tau}^2 - \underline{\tau}^2}{2} \cdot \left(v - \frac{c(\bar{\tau}^2 - \underline{\tau}^2)/(2\mu)}{\mu - (\bar{\tau}^2 - \underline{\tau}^2)/2} - \omega \right. \\ & \quad \left. - \frac{v^2(\bar{\tau}^2 + \underline{\tau}^2)}{4(\bar{\tau} + \underline{\tau})(1 - \delta)} \right) \\ & < \max_{0 \leq \tau \leq \bar{\tau} \leq 1} \frac{\bar{\tau}^2 - \tau^2}{2} \cdot \left(v - \frac{c(\bar{\tau}^2 - \tau^2)/(2\mu)}{\mu - \frac{\bar{\tau}^2 - \tau^2}{2}} - \omega \right). \quad (8) \end{aligned}$$

One can easily verify that the optimization problem (8) is equivalent to the optimal control problem (5) in the no-localization benchmark. Therefore, the optimal value of problem (8) is equal to the LLM's optimal profit Π_0^* in the no-localization benchmark, and this implies

that LLM's optimal profit Π^* with localization will be lower than Π_0^* (i.e., $\Pi^* \leq \Pi_0^*$).

Now, without localization, because the LLM provider will set the service price $p_0 = v - c(\lambda_0/\mu)/(\mu - \lambda_0)$, all cloud users will have zero utility, i.e., $u_c = 0$. As such, their surplus $US_0^* = 0 \leq US^*$. \square

The proofs of all other results are available upon request.