https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–16
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Unlocking Profits in Generative AI: The Impact of AI Adaptive Learning on Freemium Strategy

Yinliang (Ricky) Tan,[a] **Ling Zhong,**[b,*] **Michelle Y. Lu,**[c] **Jiajia Nie**[d]

[a] Decision Sciences and Management Information Systems, China Europe International Business School, Shanghai 201206, P.R. China; [b] School of Management, Fudan University, Shanghai 200433, P.R. China; [c] Marketing, China Europe International Business School, Shanghai 201206, P.R. China; [d] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, P.R. China

*Corresponding author

yrtan@ceibs.edu (Y. (R.) Tan), zhlzhongling@outlook.com (L. Zhong), michellelu@ceibs.edu (M. Y. Lu), niejiajia@home.swjtu.edu.cn (J. Nie)

https://orcid.org/0000-0001-8971-818X (Y. (R.) Tan), https://orcid.org/0009-0009-2392-8120 (L. Zhong)

Received: February 12, 2024
Revised: May 8, 2025; December 29, 2025; May 2, 2026; July 25, 2026
Accepted: September 6, 2026
Published Online in Just Accepted: September 18, 2026
Published Online in Articles in Advance: November 2, 2026

https://doi.org/10.1287/isre.2025.2140

Copyright: © 2026 INFORMS

Abstract. Although conventional software is programmed for specific tasks, artificial intelligence (AI) *learns* and *refines* its capabilities through a two-stage process: pretraining and fine-tuning. A central feature of this process is the AI adaptive learning embedded in the fine-tuning stage, whereby user interactions refine the AI's capabilities initially developed from publicly available data during pretraining. This feature reshapes the firm's freemium decision. We develop an analytical model to examine how AI adaptive learning affects firms' optimal freemium strategies and find that, although a free version can expand the user base and generate learning data, it also intensifies cannibalization of demand and increases the service-cost burden. We demonstrate that highly effective adaptive learning can paradoxically harm profitability by excessively enhancing the free version's appeal and sharply cannibalizing premium demand. We find further that firms may optimally avoid offering a free version even when service costs are minimal, relying instead on reductions to the price of premium service to expand the paid user base. Narrowing the base capability gap between versions can also increase profitability by strengthening adaptive learning despite the increased cannibalization. Finally, results show that introducing a free version does not necessarily increase consumer surplus, because improved AI capabilities may allow firms to raise premium prices. These results are robust across several model extensions and inform both firms' monetization strategies and policy discussions of consumer welfare in generative AI markets.

History: Hsing Kenneth Cheng, Senior Editor; Ling Xue, Associate Editor.
Funding: L. Zhong acknowledges support from the China Postdoctoral Science Foundation [Grant KLH1019212]. Y. (R.) Tan acknowledges financial support from the China Europe International Business School research funds [Codes AG25TRD, BU26XIA, DP25TYL, and AU26BDA].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2025.2140.

Keywords: generative AI • adaptive learning • versioning • freemium strategy

1. Introduction

Artificial intelligence (AI) has gone through repeated waves of interest, but recent advances in generative models have sharply accelerated commercial adoption. The release of ChatGPT in November 2022 and its rapid diffusion across consumer and professional contexts illustrate the commercial viability of systems capable of generating text, images, music, and computer code. Applications such as Midjourney and GitHub Copilot highlight the breadth of tasks now mediated by generative AI. As this market continues to grow (Bloomberg 2024), firms are actively experimenting with how to commercialize generative AI products.

One increasingly common approach is the freemium strategy, a specific form of software versioning in which firms offer a free version alongside a premium version (hereafter, the *freemium strategy*).[1] In traditional software markets, such versioning is primarily used to broaden market coverage, differentiate product offerings, and manage users' willingness to pay (Shivendu and Zhang 2015). In generative AI, however, the logic of freemium changes more fundamentally. Offering a free version not only expands access and segments demand but can also make broader user participation strategically valuable for improving product capability. As a result, user access may affect not only current demand, but also the development of product capability. This link between user scale and capability distinguishes generative AI freemium from standard software settings and reshapes the firm's incentive to offer multiple versions. Broader user access matters not because the mere presence of more users directly increases other users' utility, but

because user interactions generate data that the firm can internalize and use to improve the product's intrinsic capability. User scale therefore creates value through firm-mediated capability improvement rather than direct interpersonal utility. Importantly, this capability improvement is realized through the firm's learning process rather than through users' contemporaneous interaction with one another.

1.1. Motivation

Generative AI capability is built through a two-stage process: pretraining and fine-tuning (OpenAI 2023, Google 2024b). In pretraining, the model acquires broad, general-purpose competence from large corpora such as web pages and books (Achiam et al. 2024). However, this knowledge is inherently generic and may not capture the nuance required for diverse user- and context-specific tasks. Fine-tuning addresses this limitation by using user-interaction data to improve the model's performance on specific tasks, domains, and usage contexts (Achiam et al. 2024). We refer to this user-interaction-driven refinement as AI adaptive learning.[2]

Adaptive learning uses human-interaction data such as prompt-revision pairs, ratings, and pairwise preferences to improve model performance on task-specific dimensions (OpenAI 2023). Representative techniques include supervised fine-tuning and preference-based optimization (e.g., Reinforcement learning from human feedback) (Touvron et al. 2023, Google 2024a). Firms can also shape this process through their data collection, curation, and training choices, which determine how effectively user interactions are converted into capability improvement (Google 2026b, OpenAI 2026c). In practice, firms collect user feedback through their AI-based services such as ChatGPT and Gemini, including response ratings and written comments, and use such feedback to improve model performance and better align outputs with user needs (Google 2026a, OpenAI 2026b). As a result, a free version in generative AI may play a dual role: It can expand access while also generating interaction data that improves product capability.

At the same time, this broader access is not costless. Serving additional free users leads the firm to incur nontrivial and often increasing marginal service costs due to the computational demands of generating responses. Industry reports projected an 89% increase in average computing costs from 2023 to 2025 and identified generative AI as a major driver, highlighting the economic challenges firms face when managing expanded AI services (Brodsky 2024). Moreover, a stronger free version may also intensify cannibalization of premium demand. Thus, these forces create a managerial tension: the same free access that helps strengthen AI capability may also intensify cannibalization and raise substantial service costs.

This tension is reflected in industry practice. Although some AI providers offer both free and premium versions, Midjourney's main AI-based service is available only through paid subscriptions (hereafter referred to as the *paid-only strategy*). Existing research provides limited guidance on how the AI adaptive learning process interacts with freemium decisions in this context. Still, it remains unclear whether, and under what conditions, offering a free version is optimal. Addressing this gap motivates the present study.

1.2. Research Questions and Key Findings

Motivated by this setting, we examine how AI adaptive learning reshapes firms' freemium incentives and profitability in the generative AI market. We develop an analytical model to study how AI adaptive learning affects firms' optimal freemium strategies. The model captures three core forces: the learning benefit from additional user interactions, the cannibalization between free and premium versions, and the service cost burden associated with serving a larger user base. In our setting, absent adaptive learning, a zero-priced lower version only adds cannibalization and service costs; hence, the economic question is whether learning value from free users can overturn this baseline logic. Our analysis focuses on the following questions.

First, under what conditions is a freemium strategy viable? Although adaptive learning allows firms to extract value from free users by improving AI capability, we show that freemium is optimal only when learning effectiveness is neither too low nor too high. When learning effectiveness is low, data generated by free users yields limited capability improvement, making the free version largely a source of cannibalization and cost. When learning effectiveness is high, the free version may become sufficiently capable to sharply erode demand for the premium version. As a result, freemium is optimal only when learning effectiveness falls within an intermediate range.

Second, when does freemium outperform a paid-only strategy? We show that firms may optimally avoid offering a free version even when service cost burdens are minimal. In such cases, firms can instead reduce the premium price to expand the paid user base and generate sufficient learning data without incurring cannibalization from a free version. As service costs rise to moderate levels, however, aggressive price reductions for the premium version become less attractive, and introducing a free version becomes a more effective way to expand the learning data set and improve AI capability. Freemium is therefore optimal only over specific ranges of the service-cost coefficient.

Third, how should firms differentiate between free and premium versions? Contrary to conventional wisdom, we show that narrowing the base model capability gap between versions can increase profitability. A

moderately capable free version expands the user base and strengthens adaptive learning, raising the value of the premium version sufficiently to offset increased cannibalization.

Finally, we examine consumer welfare. Although free access may appear beneficial, we show that introducing a free version does not necessarily increase consumer surplus. Improvements in AI capability may enable firms to raise premium prices, allowing them to capture much of the surplus generated by adaptive learning.

1.3. Literature Review

This study relates to two main streams of literature: research on AI and human-AI interaction and research on software versioning.

First, our work contributes to the growing literature on AI in operations, strategy, and information systems (Pei et al. 2025). AI typically includes two attributes, *automation* (Berente et al. 2021) and *smartness* (Raisch and Krakowski 2021), which enable the AI to be applied to automating processes and improving human work outcomes through its capacity to make smart decisions (Hou et al. 2024, Gao et al. 2025). In one related study, Li and Li (2022) explore the role of AI automation in assisting a regretful retailer to address the newsvendor problem, whereas Cui et al. (2022) consider how AI creates and delivers value to buyers in the context of procurement problems. Related studies analyze AI-enabled demand forecasting, procurement, and knowledge contribution (Ferreira et al. 2016, Shan and Qiu 2026).

The above literature, however, is mainly concerned with how market players leverage AI's off-the-shelf automation and smartness capabilities, leaving unexplored the internal mechanisms driving AI performance improvements. We complement this work by explicitly modeling the learning process through which AI capability improves. Although Gurkan and de Véricourt (2022) consider a similar machine-learning process, their analysis typically integrates training stages without explicitly recognizing their unique contributions. By distinguishing between pretraining and fine-tuning, we highlight how user interaction data contributes to adaptive learning and shapes firms' strategic incentives. Unlike Gurkan and de Véricourt (2022), who abstract from learning stages or assume homogeneous preferences, our framework introduces heterogeneous consumer preferences, where different users have distinct expectations at different training stages. This heterogeneity reduces the firm's incentive to acquire data compared with homogeneous preferences. From a strategic perspective, our research highlights the strategic management of freemium to balance AI adaptive learning benefit, cannibalization risks, and service cost burden, whereas Gurkan and de Véricourt (2022) center on pricing and contracting strategies within the framework of the AI learning, highlighting a more contractual and pricing-based approach. Therefore, our analysis bridges these AI learning dynamics with product freemium strategy, offering insights into how firms can leverage AI adaptive learning to optimize AI capabilities, manage cannibalization risks, and control service cost burden.

Second, the freemium strategy examined in our study is a specific form of software versioning; our work therefore also contributes to the software versioning literature. Existing research examines how firms design product lines under heterogeneous preferences (Shivendu and Zhang 2015), outside options (Chen and Seshadri 2007), piracy risk (Chellappa and Shivendu 2005, Wu and Chen 2008), and network effects (Cheng and Liu 2012). In these studies, the incentive to offer multiple versions is driven mainly by demand segmentation, market expansion, and cannibalization management. By contrast, in our setting, a free version can expand the user base and generate learning value, thereby improving product capability through adaptive learning.

Although the user base also plays a role in the network-effect settings examined in the existing versioning literature, our AI adaptive learning mechanism is fundamentally different. In standard network-effect settings, a larger user base directly raises users' valuation because the product becomes more valuable when more users are involved. In our setting, by contrast, additional users matter because their interactions generate data that the firm can internalize and use to improve the capability of the AI product. Moreover, this learning process is mediated by the firm: The value of user interactions depends on the firm's ability to collect, curate, and train on those data. Users therefore benefit not from interpersonal utility generated by the presence of other users, but from firm-mediated improvement in intrinsic product capability.

This difference further leads to a different versioning logic. Stronger user interaction does not necessarily make versioning more attractive, because the same learning process can also strengthen the capability of the free version and erode differentiation. Accordingly, the profitability of versioning depends on how learning gains interact with cannibalization. This differs from the standard network-effects logic, under which versioning can remain attractive even when network effects are relatively weak (Cheng and Liu 2012). This different logic also yields a distinct implication: reducing the base-model capability gap can, counterintuitively, increase firm profitability when the adaptive-learning benefit more than offsets the resulting loss in differentiation, thereby extending the existing versioning literature (Chen and Seshadri 2007, Wei and Nault 2014). More broadly, unlike standard network-effect formulations, our framework allows the value of additional user interactions to diminish as the user base grows, which better reflects practical constraints in AI training and deployment (Knight 2023).

Our work also contributes to this literature by explicitly addressing the critical role of marginal costs, an aspect largely overlooked in prior studies, with a few notable exceptions (Bhargava and Choudhary 2008, Chellappa and Mehra 2018). Although these studies examine how marginal cost affects a monopolist's versioning decision, they assume that cost increases linearly with the number of users. This assumption is less suitable in generative AI settings, where serving additional users often requires substantial and increasingly costly computation. To address this, our research expands the scope of the existing literature by incorporating a quadratic cost form to model the unique cost structure of AI. This extension contributes to the literature in two ways. First, our findings complement Bhargava and Choudhary (2008) by demonstrating that the firm may refrain from providing a low-quality free version even when the variable costs are minimal. Second, our findings extend Chellappa and Mehra (2018) by showing that, even under strictly increasing marginal costs typical of AI applications, freemium may still be optimal. This is because the adaptive learning benefits generated by additional users can outweigh the associated cost increase. To the best of our knowledge, this represents the first comprehensive analytical study to explore increasing marginal costs within versioning strategies explicitly tailored to the context of generative AI.

Taken together, these differences show that AI adaptive learning changes the underlying incentive logic of versioning rather than merely adding another benefit to a conventional framework. In our setting, a free version is valuable not only as a segmentation approach but also as a learning instrument. The incentive to offer multiple versions therefore depends on whether the learning gains from broader user access are strong enough to offset the associated loss in cannibalization and the service-cost burden.

The remainder of the study is organized as follows. Section 2 introduces the base model. We present the analysis and results in Section 3. In Section 4, we extend our base model along several different directions. We conclude this study and provide managerial insights in Section 5. All proofs and notations are provided in the Online Appendix.

2. Model

We consider a vertically integrated generative AI provider (hereafter, the firm) that develops the underlying foundation model and commercializes it through an AI-based service. The firm chooses between two strategies in its versioning decision: offering only a premium version (the *paid-only strategy*) or offering both a premium and a free version (the *freemium strategy*). Consumers decide whether to adopt the service and, when applicable, which version to use.

The model captures a core feature of generative AI: Its capability evolves through a two-stage learning process of pretraining and fine-tuning. We focus on how user interactions during fine-tuning contribute to adaptive learning and how this learning process shapes the firm's pricing and freemium decision.

2.1. AI

Our analytical model, although an abstraction, maps directly to the two-stage learning process observed in the generative AI industry. The focal feature of generative AI is its adaptive learning process during the fine-tuning stage, which lies at the core of our model. In general, AI capability is developed through pretraining and fine-tuning stages (Raffel et al. 2020). The pretraining stage establishes the model's fundamental capability through training on publicly available sources (Google 2024b). We define the AI capability developed during this stage as the base model capability θ^B. Without loss of generality, we assume θ^B is exogenous in the base model. In the model extension, we relax this assumption and endogenize θ^B, with the main qualitative insights remaining robust.

Building on the base model, the AI improves its performance in the fine-tuning stage by learning from curated data through an adaptive learning process (Touvron et al. 2023Brown et al. 2020). Following practices akin to Gemini, the curated human-interaction data considered in our setting include prompt-revision pairs, explicit ratings, and pairwise preferences. Through supervised fine-tuning and preference-based optimization, the model is trained on these data to improve factual accuracy, stylistic alignment, and performance on user-specific tasks beyond its general-purpose capabilities (Google 2024b).

We introduce θ^I as the AI adaptive learning increment, representing the improvement in AI capability during the fine-tuning stage, and specify it as

$$\theta^I = \alpha \cdot \theta^B \cdot \Delta N. \tag{1}$$

Here, $\alpha \in (0,1)$ measures how effectively the AI model can transform additional user-interaction data into capability improvements through AI adaptive learning. Operationally, α is shaped by the firm's data infrastructure, curation processes, training pipeline, optimization techniques, and organizational capabilities (Google 2026b). Thus, the firm can determine the extent to which user-interaction data can be internalized and converted into improved model capability. A higher α therefore reflects a stronger ability to extract value from user feedback and translate it into model improvement. The term θ^B is the base model capability that serves as the basis of fine-tuning, and ΔN is the total number of users, reflecting the practice of AI providers such as OpenAI and Google of collecting user-generated data to enhance their models. We allow the adaptive learning increment θ^I to depend

on the base model capability θ^B, capturing the idea that models with stronger baseline capabilities are better able to leverage fine-tuning for task-specific adaptation.

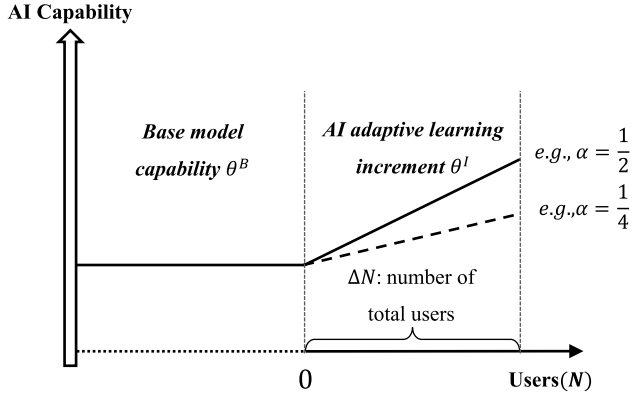
This formulation also clarifies the role of the user base in our model. Unlike in a standard network-effect specification, user base ΔN does not directly increase user utility. Instead, it affects utility only indirectly through θ^I, by enlarging the size of user-interaction data from which the AI learns. The magnitude of this effect is governed by α, which reflects how effectively the firm can internalize and use those data. Thus, the impact of additional users is mediated by the firm's learning capability rather than arising mechanically from user presence itself. The learning term should therefore be interpreted as the expected capability increment generated through the firm's fine-tuning process from the equilibrium user base. Even though the baseline model is single period, it can still be interpreted as a reduced-form representation in which consumers take this firm-mediated capability improvement into account when making adoption decisions; the user base itself does not generate contemporaneous interpersonal utility as in a network-effect model. Section 4.1 further makes this timing structure explicit in a two-stage game with forward-looking consumers and shows that the main qualitative insights continue to hold.

Accordingly, aggregate AI capability is given by the sum of the base model capability and the adaptive learning increment, which is illustrated in Figure 1.[3]

$$\theta^B + \theta^I = \theta^B + \alpha \cdot \theta^B \cdot \Delta N \tag{2}$$

Note that to obtain closed-form solutions, we assume the adaptive learning increment is linear in the number of users in the main model. In the extension section, we consider an alternative specification in which the adaptive learning increment is concave in the number of users, that is, $\theta^I = \alpha \cdot \theta^B \cdot \sqrt{\Delta N}$. This concave functional form captures the diminishing marginal return to additional users and allows us to illustrate numerically that our main qualitative results remain unchanged.

Figure 1. AI Learning Process



2.2. Firm

In keeping with industry practice, we consider that the firm may offer two versions of an AI-based service: a premium version (P) and a free version (F), with the premium version delivering higher functionality. Differences between the two versions arise primarily from the pretraining stage, as more advanced AI models are typically trained with a larger number of parameters and more extensive datasets (Maslej et al. 2025).[4] We therefore capture version differentiation through differences in base model capability, denoted by θ^B.

The firm chooses between two strategies: offering only the premium version (the *paid-only strategy*, denoted by superscript S) or offering both the premium and free versions (the *freemium strategy*, denoted by superscript V). Under the paid-only strategy, the firm charges a subscription price p for the premium AI service (Feng et al. 2018). Under the freemium strategy, the firm sells the premium version while also making the free version available.

Increased service costs stemming from the computational resources required to respond to user inputs and queries will be incurred; as the volume of consumer-generated prompts rises, the service cost burden associated with the billions of calculations powering the AI's effective responses will increase as well (Loten 2023, Vanian and Leswing 2023). Recent research documents increased marginal costs for AI operations (Gurkan and de Véricourt 2022), so we assume accordingly that the service cost takes the form of $s\,(\Delta N)^2/2$, where $s \in (0,1)$ denotes the firm's service cost coefficient and ΔN the total number of users. We relax this assumption in the extension to allow the service cost coefficients to differ across versions. Following Chellappa and Mehra (2018), we assume that AI development costs are sunk.

2.3. Consumers

Consumers are heterogeneous in their willingness to pay and in their valuation v of the base model capability θ^B, where v is uniformly distributed on [0, 1] (Cheng et al. 2015, Feng et al. 2019, Tan 2024). A consumer of type v derives utility $v \cdot \theta^B$ from the base model capability.[5] This formulation captures heterogeneity arising from differences in efficiency requirements, task complexity and privacy sensitivity. For a given θ^B, a more privacy-sensitive user (preferring outputs generated with fewer prompts and thus a lower probability of privacy leakage) or a user with more complex tasks (requiring broader and more accurate functionality to avoid external tools) assigns a lower effective valuation; conversely, users with lower privacy sensitivity or simpler tasks assign a higher valuation. Because of the adaptive learning nature of AI, both free and premium users benefit from the learning increment θ^I, because the AI model improves through user interaction during the fine-tuning stage (Brown et al. 2020).[6] In the extension, we also allow consumers to have heterogeneous preferences for both

θ^B and θ^I. We normalize the mass of potential consumers to one.

Consistent with the literature, we assume that consumers face version-specific outside options when deciding whether to adopt the AI service (Chen and Seshadri 2007). We model the outside options as version-specific participation thresholds induced by available market alternatives.[7] These outside options apply uniformly across consumers for each version. Specifically, we normalize the outside option for the premium version to zero and assume the outside option for the free version is c. This reflects the fact that the premium version, with superior functionality, typically faces fewer effective substitutes. We then extend this model to allow the premium outside option to be either higher or lower than that of the free version. The outside option c is a critical component of the freemium model (Cheng and Liu 2012), because ignoring it would render demand for the free version independent of AI capability.

Without loss of generality, we let $\theta^B = 1$ for the premium version and $\theta^B = q$ for the free version, where $1 - q \in (0,1)$ measures the gap in base model capability between the two versions. Let N_P and N_F denote the total number of premium and free users, respectively. In the base model, we adopt a single-period framework in which the firm and consumers make decisions at the beginning of the fine-tuning stage, focusing on the core adaptive learning mechanism. In the extension, we analyze a two-period game in which decisions are made at the beginning of the pretraining stage and agents form rational expectations about future learning gains.

If the firm adopts the freemium strategy, consumers choose between the premium and free versions, with consumers of the premium version deriving utility

$$U_P^V(v) = v \cdot \theta^B + \alpha \cdot \theta^B \cdot \Delta N - p = \underbrace{v}_{\substack{\text{Perceived}\\ \text{base model}\\ \text{capability}}} + \underbrace{\alpha(N_P + N_F)}_{\substack{\text{Perceived AI}\\ \text{adaptive}\\ \text{learning}\\ \text{increment}}} - \underbrace{p}_{\substack{\text{Price of the}\\ \text{premium}\\ \text{version}}},$$

and consumers of the free version deriving utility

$$U_F^V(v) = v \cdot \theta^B + \alpha \cdot \theta^B \cdot \Delta N - c = \underbrace{qv}_{\substack{\text{Perceived}\\ \text{base model}\\ \text{capability}}} + \underbrace{q\alpha(N_P + N_F)}_{\substack{\text{Perceived AI}\\ \text{adaptive}\\ \text{learning}\\ \text{increment}}} - \underbrace{c}_{\substack{\text{Outside}\\ \text{options}}}.$$

If the firm adopts the paid-only strategy, then consumers decide whether to adopt the premium version, with premium consumers deriving utility

$$U_P^S(v) = v \cdot \theta^B + \alpha \cdot \theta^B \cdot \Delta N - p = v + \alpha N_P - p.$$

The timeline of the game is as follows. First, the firm chooses between the freemium and paid-only strategies. Then, it sets a premium price p. Finally, consumers make their adoption decisions. The notations used in this study are summarized in Table 1.

In the main text, we focus on the parameter region that jointly ensures partial market coverage and rules out the trivial case in which the free version never emerges. Specifically, we impose $s > \max\{0, 2\alpha - 1\}$ and $q\alpha < c < q$. Nevertheless, the complementary cases, including $0 < s \le \max\{0, 2\alpha - 1\}$ and $c \le q\alpha$, are discussed in the Online Appendix. We emphasize this baseline region because it is the economically and managerially relevant setting in which a freemium strategy can expand the total user base and thereby strengthen adaptive learning, while still exposing the firm to cannibalization and service cost tradeoffs. By contrast, if the market is already fully covered, introducing a free version cannot attract additional users, so the learning benefit from user-base expansion no longer operates. If the free version does not arise in equilibrium, the choice between the freemium and paid-only strategies becomes trivial. Both cases therefore move our base model away from the central mechanism that motivates our analysis.

3. Results

In this section, we first analyze equilibrium pricing and demand under the freemium and paid-only strategies to examine how AI adaptive learning affects these

Table 1. Parameters and Decision Variables

Symbol	Definition
θ^B	Base model capability developed in the pretraining stage, that is, $\theta^B = 1$ for the premium version; $\theta^B = q$ for the free version, $q \in (0,1)$
θ^I	AI adaptive learning increment in the fine-tuning stage
ΔN	The total number of users, or the learning data set size of the fine-tuning stage
α	AI adaptive learning effectiveness, $\alpha \in (0,1)$
s	Service cost coefficient for responding to user inputs, $s \in (\max\{0, 2\alpha - 1\}, 1)$
p	Price for the premium version
v	Consumer heterogeneous valuation for the base model capability, $v \sim U(0,1)$
c	Consumer outside option for the free version, $c \in (q\alpha, q)$
N_P	Total number of premium users
N_F	Total number of free users
B	This superscript indicates the base model capability
I	This superscript indicates the AI adaptive learning increment
P	This subscript indicates the premium version
F	This subscript indicates the free version
S	This superscript indicates the case of the paid-only strategy
V	This superscript indicates the case of the freemium strategy
π	Firm profit
cs	Total consumer surplus

outcomes. We then study the firm's optimal freemium decision and characterize how the learning process shapes this choice. Finally, we analyze the impact of introducing a free AI version on consumer surplus.

3.1. Freemium Strategy

Under the freemium strategy, both the premium and the free versions are available. The firm sets a premium price p, with profit given by

$$\pi^V(p) = pN_P - \frac{1}{2}s(N_P + N_F)^2,$$
$$\text{s.t. } N_P > 0,\ N_F > 0.$$

The following lemma summarizes the equilibrium price, demand, profit and consumer surplus under the freemium strategy.

Lemma 1. *Under the freemium strategy, the firm's equilibrium price, demand, profit, and consumer surplus are as follows:*

(i) *If* $q\alpha < c \le \frac{q(1-q)}{2-q-\alpha}$ *and* $\max\{0, 2\alpha - 1\} < s \le \min\left\{\frac{(\alpha c - cq + q^2 - q)^2}{2(1-q)(q-c)^2}, 1\right\}$,

(a) $p^{V*} = \frac{(1+c)q - q^2 - c\alpha}{2q(1-\alpha)}$;

(b) $N_P^{V*} = \frac{1}{2q}\left[\frac{q-c}{1-\alpha} + \frac{c}{1-q}\right]$, $N_F^{V*} = \frac{1}{2q}\left[\frac{q-c}{1-\alpha} - \frac{c}{1-q}\right]$;

(c) $\pi^{V*} = \frac{1}{4q^2}\left[\frac{c^2}{1-q} + \frac{2c(q-c)}{1-\alpha} + \frac{(q-c)^2(1-q-2s)}{(1-\alpha)^2}\right]$;

(d) $cs^{V*} = \frac{1}{8q^2}\left[\frac{c^2}{1-q} + \frac{2c(q-c)}{1-\alpha} + \frac{(q-c)^2(1+3q)}{(1-\alpha)^2}\right]$;

(ii) *Otherwise, the freemium strategy is not feasible.*

Lemma 1 shows that the freemium strategy is feasible only when the free version attracts a nontrivial user base and service costs are not prohibitively high. This requires relatively low outside options for the free version, so that a zero-priced but basic AI service is attractive, and a moderate service-cost coefficient, so that the service cost of responding to additional user interactions does not overwhelm the learning benefits. Under these conditions, the firm balances three forces when setting the premium price: cannibalization from the free version, the positive learning effect from a larger total user base, and service cost burden. The equilibrium outcomes in Lemma 1 serve as the benchmark for comparison with the paid-only strategy.

3.2. Paid-Only Strategy

Under the paid-only strategy, the firm only offers the premium version to consumers and decides the premium price p. Profit is derived as

$$\pi^S(p) = pN_P - \frac{1}{2}s{N_P}^2,$$
$$\text{s.t. } N_P > 0.$$

Lemma 2. *Under the paid-only strategy, for all* $q\alpha < c < q$ *and* $\max\{0, 2\alpha - 1\} < s < 1$, *the equilibria are*

$$p^{S*} = \frac{1-\alpha+s}{(2-2\alpha+s)},\ N_P^{S*} = \frac{1}{(2-2\alpha+s)},$$
$$\pi^{S*} = \frac{1}{2(2-2\alpha+s)},\ cs^{S*} = \frac{1}{2(2-2\alpha+s)^2}.$$

Lemma 2 provides the benchmark under which the firm avoids both cannibalization and the added service-cost burden of free users. Although adaptive learning still occurs, it is driven solely by premium users, which may result in a smaller learning data set than under the freemium strategy.

3.3. Firm's Freemium Decision

We now examine when a firm adopts the freemium strategy and when it opts for a paid-only strategy.

Additional free users generate three effects. First, a learning effect: Users' interactions (prompts, revisions, ratings) expand the training data set used in AI adaptive learning, which increases the AI's capability increment and thus improves the capability of both the free and the premium versions. This raises users' valuations and, through higher willingness to pay, can increase optimal premium pricing. Second, a cannibalization effect: Some free users would otherwise have purchased the premium version, so attracting them to the free version instead reduces premium demand and revenues. Third, an increasing marginal service-cost burden: serving the additional user requires computational resources, and with a convex service-cost function, the marginal cost of serving each extra user rises with the total user base.

The net value of free users equals the learning benefit minus cannibalization and service-cost burden. Learning strength depends on adaptive learning effectiveness α, reflecting how efficiently the firm can convert extra interaction data into capability improvements; cannibalization depends on the base model capability gap and outside options; and service costs depend on s and total usage, which together determine how quickly marginal costs escalate as the user base grows.

Proposition 1. *For the freemium strategy to be optimal, the free version's base model capability* q *and adaptive learning effectiveness* α *must satisfy*[8]

$$0 < q < \frac{1}{2} \text{ and } q < \alpha < 1 - q.$$

Proposition 1 identifies a "Goldilocks" region, where $0 < q < 1/2$ ensures the nonemptiness of this region. When adaptive learning is weak, free users generate little learning value and mainly create cannibalization and cost. When learning is very strong, the free version becomes too capable relative to the premium version, intensifying cannibalization. The freemium strategy is

viable only when learning is effective, but differentiation remains meaningful.

This result highlights the key economic implication of adaptive learning in our setting: The value of broader adoption depends not simply on attracting more users, but on whether the resulting user interactions generate enough capability improvement to outweigh the associated loss in differentiation.

From a managerial perspective, if α is low, firms should invest in learning infrastructure before relying on free access. If α is high, the freemium strategy can be profitable only if the free version's base model capability is deliberately constrained. Otherwise, the learning effect will make the free version "too good," triggering severe cannibalization even before service-cost burdens bind.

Proposition 1 identifies when the freemium strategy is structurally viable, requiring sufficient learning value without eliminating differentiation between the two versions. Conditional on this region, Proposition 2 shows how the outside option c and service-cost coefficient s determine whether the learning gains from broader access outweigh cannibalization and service costs.

Proposition 2. *Comparing profits under the freemium and paid-only strategy, $\pi^{V*} > \pi^{S*}$ if either*

(i) *The parameters satisfy $q\alpha < c \le c_1$ and $\max\{0, 2\alpha - 1\} < s < s_2$, or*

(ii) *The parameters satisfy $c_1 < c \le \frac{q(1-q)}{2-q-\alpha}$ and $s_1 < s < s_2$.*

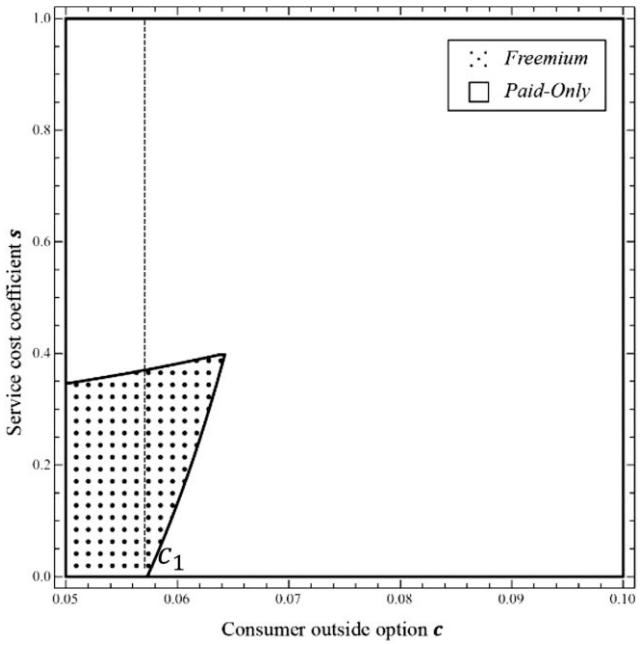
Otherwise, $\pi^{V*} < \pi^{S*}$.

The variables c_1, s_1, and s_2 are defined in the Online Appendix. The shaded area in Figure 2 illustrates the conditions under which the freemium strategy dominates. Conventional wisdom may suggest that firms would consistently offer a free version when service costs are low (i.e., s is low). However, Proposition 2(ii) demonstrates this intuition may not hold, particularly when the consumer's outside options c are moderate. The underlying reasons are as follows.

The outside option c shapes how many users would adopt the free version at a zero price. When c is high, few consumers are willing to use the free version even if it is offered, so it contributes little additional data for adaptive learning while still posing cannibalization risks and cost burden. In this case, the freemium strategy is never profitable, and the firm optimally offers only the premium version.

When c is low or moderate (i.e., $q\alpha < c \le c_1$), the tradeoff becomes more nuanced. With low c, a free version attracts a large user base and therefore a rich learning data set but also brings severe cannibalization effects and generates substantial service costs. The freemium strategy is then attractive only when service cost burdens are sufficiently low so that the additional learning value from free users outweighs both service costs

Figure 2. Optimal AI Freemium Decision Under Linear $\boldsymbol{\theta}^I$ Scenario



Notes. The figure depicts the firm's optimal strategy as a function of the service cost coefficient and outside options. Parameters are set to $q = 0.1$ and $\alpha = 0.5$, with the qualitative patterns robust to a wide range of alternative settings.

and cannibalization losses. Under moderate c (i.e., $c_1 < c \le q(1-q)/(2-q-\alpha)$), the free version would attract a smaller but still meaningful number of users. If service cost burdens are very low, the firm can instead lower the premium price to build a large paid user base, obtain ample learning data, and avoid cannibalization from a free version, so the single-version strategy dominates. As service costs rise to a moderate level, however, aggressive premium price cuts become too expensive, limiting the size of the paid user base and the resulting learning data set. In this region, introducing a free version is profitable the adaptive learning substantially strengthens AI capabilities, thereby raising the premium version's value enough to offset cannibalization and service costs. Thus, the freemium strategy is optimal specifically at moderate, not minimal, service costs under moderate c.

Proposition 2 provides a decision rule for managers regarding whether to adopt the freemium strategy based on how the three forces change with the market environment. The most actionable and nonobvious case is moderate outside options: If service costs are very low, managers should typically expand learning via a larger paid base (price cuts) and retain the paid-only strategy to avoid cannibalization; as service costs rise to a moderate range, deep premium price cuts become inefficient, and a free version becomes the cheaper way to expand the data set, making the freemium strategy profitable despite

some cannibalization. Conversely, at high consumer outside options, firms are better off entirely avoiding offering a free version, as the potential gains from AI adaptive learning are too limited to justify any associated cannibalization risks.

Proposition 3. *Under the freemium strategy, as the free version's base model capability q increases,*

(i) *Both N_P^{V*} and N_F^{V*} increase;*

(ii) *The premium price p^{V*} first increases then decreases;*

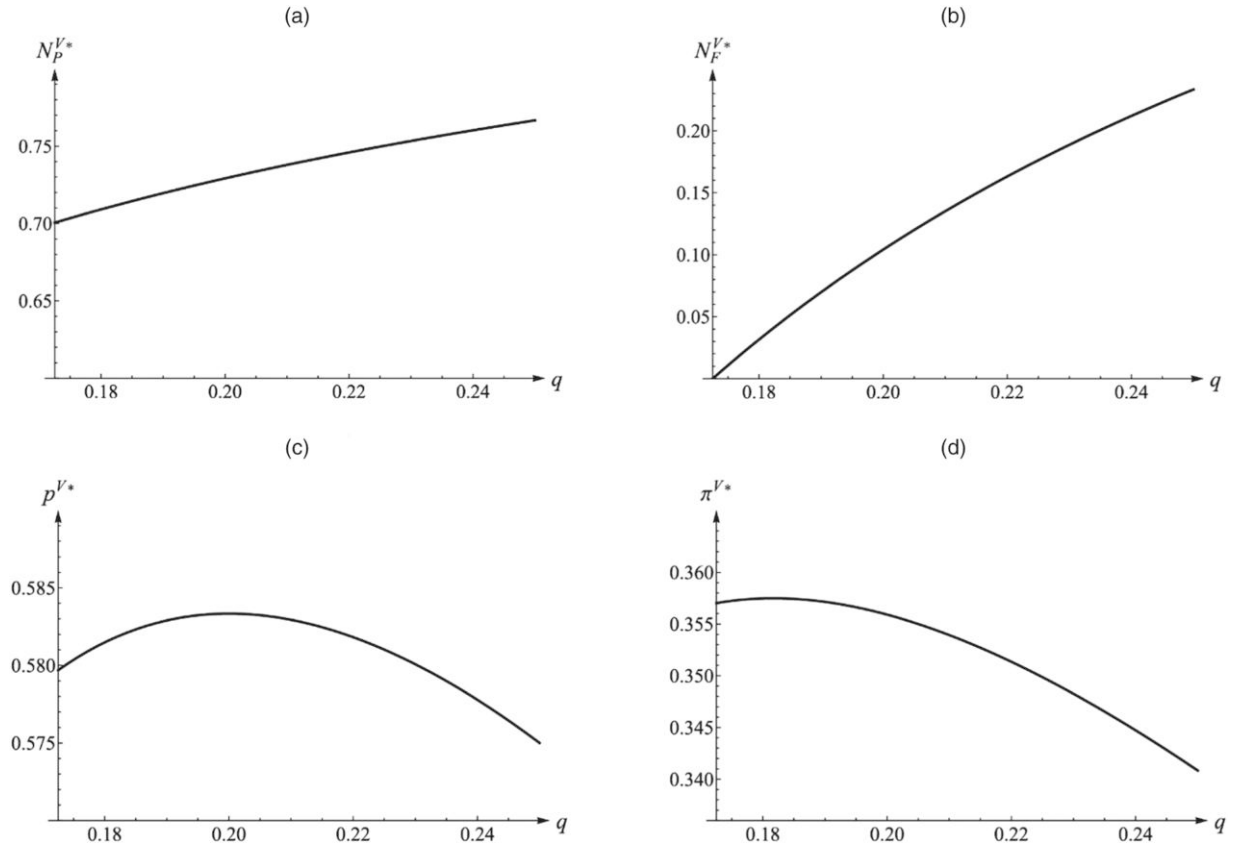
(iii) *Firm profit π^{V*} first increases then decreases if $\tilde{s}(q_H) < s < \tilde{s}(q_L)$.*

The variables $\tilde{s}(q_H)$ and $\tilde{s}(q_L)$ are defined in the Online Appendix. Proposition 3 analyzes how the firm should choose the base model capability gap between the free and premium versions once it has committed to a freemium strategy. In our model, this gap is governed by the free version's base model capability q. As q increases, the capability gap between the two versions shrinks, potentially intensifying cannibalization. A natural intuition is that a stronger free version should always hurt the firm by attracting users away from the premium version and, consequently, a drop in the firm's profit. However, Proposition 3 suggests that both the price and the profit may actually increase with q. Figure 3 graphically illustrates Proposition 3, and the underlying rationale of this result is as follows.

As q increases, the free version becomes more attractive, which expands the total user base and strengthens the adaptive-learning benefit. At the same time, however, the smaller capability gap intensifies cannibalization from the premium version to the free version. When q is initially low, the adaptive-learning benefits dominate: a stronger free version brings in more users, generates more learning data, and improves the overall capability of the products, allowing the firm to charge a higher premium price and earn higher profit. But when q becomes sufficiently high, the cannibalization effect becomes dominant: The free version becomes too close to the premium version, cannibalization intensifies, and the firm is forced to lower the premium price, which eventually reduces profit. This generates the inverted-U pattern: Increasing the free version's capability initially raises the premium price and the firm's profit, but beyond a threshold, further narrowing the capability gap reduces profitability even as both versions' user bases continue to grow.

Proposition 3 therefore shows that adaptive learning changes the traditional versioning tradeoff. In the

Figure 3. Impact of the Free Version's Base Model Capability



Notes. The figure depicts how equilibrium outcomes vary with the free version's base capability. Parameters are set to $c = 0.1$, $\alpha = 0.4$ and $s = 0.2$, with the qualitative patterns robust to alternative values.

traditional view, the lower version should remain sufficiently weak so that cannibalization is limited and can be covered by the additional revenue generated from lower-end users (Wei and Nault 2014). In our setting, however, a stronger free version can also expand the user base and strengthen adaptive learning. As a result, the optimal versioning design need not maximize vertical differentiation. Instead, the firm may prefer an intermediate capability gap, under which the gains from user-base expansion and adaptive learning are strong enough to outweigh the loss from cannibalization. Practically, managers should not think of free version design as "the weaker the better." Concretely, the firm should tune the free version's foundation model to sit in the "moderate gap" region where learning gains still outweigh cannibalization rather than maximizing vertical differentiation.

We next compare consumer surplus across strategies.

Proposition 4. *Comparing consumer surplus under the freemium strategy* (cs^{V*}) *and the paid-only strategy* (cs^{S*}),

(i) *If* $q\alpha < c \le \frac{q(1-q)}{2-q-\alpha}$ *and* $s_3 < s \le min\left\{\frac{(\alpha c - cq + q^2 - q)^2}{2(1-q)(q-c)^2}, 1\right\}$, *then consumers benefit from the introduction of the free version,* $cs^{V*} > cs^{S*}$;

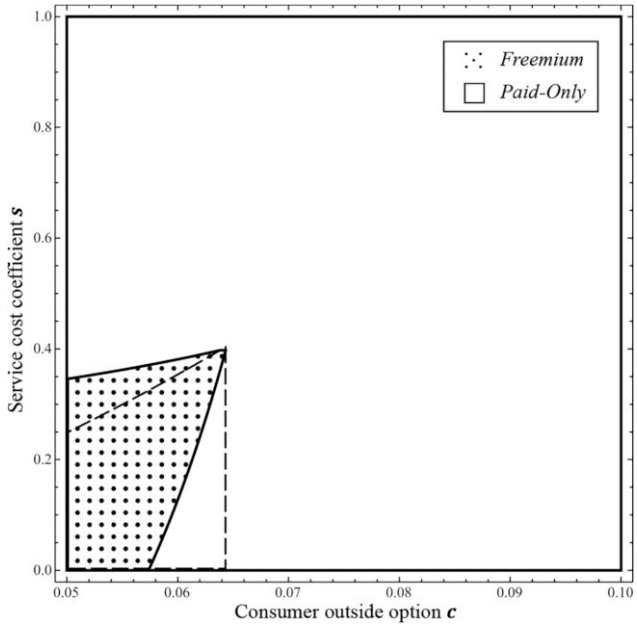
(ii) *Otherwise, consumers suffer from the introduction of the free version,* $cs^{V*} < cs^{S*}$.

Proposition 4 compares the total consumer surplus between strategies, where s_3 is defined in the Online Appendix. The proposition shows that introducing a free version does not necessarily increase consumer surplus. Consumers benefit from the free version only when the service-cost coefficient is sufficiently high.

The result follows from three forces. First, the free version expands the total user base, increasing interaction data and strengthening adaptive learning. This raises AI capability for both versions and increases user utility (holding prices fixed), which is positive for consumer surplus. Second, the free version attracts some users who would otherwise rely on outside options, generating an access effect that further increases consumer surplus. Third, however, the stronger product capability generated by adaptive learning increases users' willingness to pay for the premium version, which gives the firm greater pricing power and allows it to appropriate part of the value created. This pricing effect reduces surplus for premium users.

The comparison across strategies therefore depends on which of these three forces dominates. To connect this consumer surplus comparison with the firm's freemium decision, Figure 4 jointly depicts the regions identified in Propositions 2 and 4. The dotted region indicates where the freemium strategy generates higher firm profit than the paid-only strategy, whereas the region enclosed by the dashed outline indicates where it yields lower consumer surplus. Comparing these two regions reveals whether the firm's incentive to offer a free version is aligned with consumer welfare.

Figure 4. Impact of the Free Version on Consumer Surplus



Notes. The dotted region indicates higher firm profit under the freemium strategy, whereas the dashed outline encloses the region of lower consumer surplus under the freemium strategy. Parameters are set to $q = 0.1$ and $\alpha = 0.5$, with qualitative patterns robust to alternative values.

When the service-cost coefficient is low, the outside option determines how the firm-profit and consumer-surplus effects diverge. When the outside option is low, the firm has a strong incentive to expand the free user base aggressively. This increases the learning benefit and the access benefit, but it also substantially strengthens the firm's ability to raise the premium price. In this region, the pricing effect can dominate, so total consumer surplus under the freemium strategy is lower despite the availability of a free version. This outcome is represented by the portion of the dotted region inside the red outline in Figure 4: The freemium strategy is profitable for the firm but reduces consumer surplus, revealing a divergence between private incentives and consumer welfare.

When the service-cost coefficient remains low but the outside option is relatively high, the free version attracts fewer users. This weakens adaptive learning and limits the firm's ability to raise the premium price. The resulting revenue gain is therefore insufficient to make the freemium strategy more profitable than the paid-only strategy. Consumers also suffer in this region despite the availability of a free version. Thus, the freemium strategy yields both lower firm profit and lower consumer surplus than the paid-only strategy. This outcome occurs inside the dashed outline but outside the dotted region in Figure 4.

When service costs rise (but the freemium strategy remains feasible), the firm expands the free user base more cautiously, limiting both learning gains and the scope for premium price increases; in this region, consumer surplus is more likely to be higher under the freemium strategy. The portion of the dotted area above the sloping dashed boundary represents a win-win outcome in which free version introduction increases both firm profit and consumer surplus.

Proposition 4 thus highlights a counterintuitive welfare implication of adaptive learning–based versioning: Free access does not necessarily improve consumer welfare. Even when the free version expands access and improves product capability, the firm may capture a sufficiently large share of the resulting value through pricing, leaving consumers worse off overall.

4. Extensions

We adopt some stylized features in the base model to gain sharp insights. In this section, we extend our base model in several directions. First, we extend our model to a two-stage game. Second, we investigate the case where the firm can optimally design the base model capability. Third, we consider the scenario where the AI adaptive learning increment is concave with the learning data set size. Fourth, we analyze the case where consumers have heterogeneous preferences for both the base model capability and the AI adaptive learning increment. Finally, we allow different outside options and heterogeneous service costs across versions separately. To keep the main text focused, we report only the central result and intuition for each extension here, while detailed derivations and additional numerical results are provided in the Online Appendix.

4.1. Two-Stage Game

In the baseline analysis, we used a single-period model in which the firm and consumers make decisions at the beginning of the fine-tuning stage. We now extend the model to a two-period framework that more directly reflects the two-stage learning process, allowing the firm's strategy and consumers' expectations to evolve over time.

In the two-stage game, the pretraining stage represents the current period. The firm chooses between the freemium and paid-only strategies, and consumers make adoption decisions based on both current value from the base model and expected future value from capability improvements realized in the fine-tuning stage.

The two-stage setting preserves the baseline comparative statics: Higher AI adaptive learning effectiveness tends to reduce firm profits, whereas narrowing the capability gap between the free and premium versions may increase profitability. We next examine how forward-looking behavior alters the firm's freemium decision.

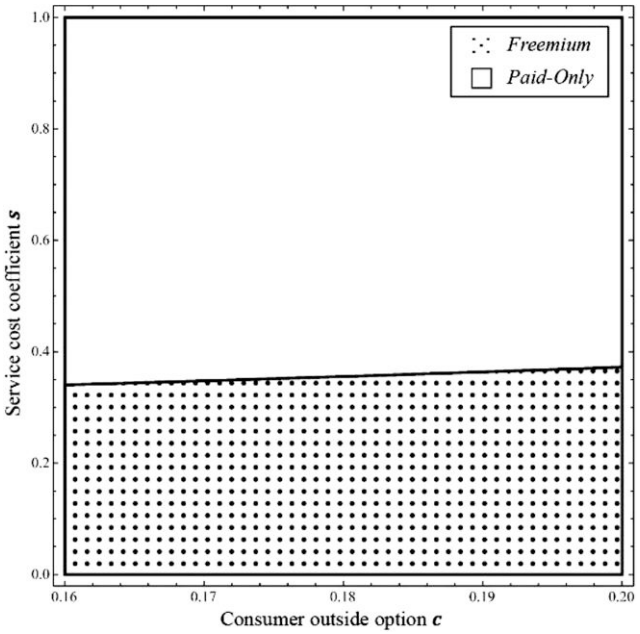
Proposition 5. *In a two-stage game, comparing the firm's profits under the freemium* (π^{V*}) *and paid-only strategy* (π^{S*}),

(i) *If* $2q < \alpha < 1$, $q\alpha < c \le q$ *and* $0 < s < s_4$, *the firm prefers the freemium strategy,* $\pi^{V*} > \pi^{S*}$;

(ii) *Otherwise, the firm prefers the paid-only strategy,* $\pi^{V*} < \pi^{S*}$.

Here, s_4 is defined in the Online Appendix. Proposition 5 reverses the base-model result when c is moderate, where the paid-only strategy dominates. In the base model, low service cost makes it attractive for the firm to lower the premium price, expand the paid user base, and obtain learning data without incurring cannibalization from a free version. In the two-stage game, by contrast, consumers anticipate future capability improvement. This makes a free version more attractive as a tool for building a larger interaction data set early, which raises expected future capability and, in turn, current willingness to pay for the premium version. When service cost is low enough (e.g., $0 < s < s_4$), this forward-looking learning effect can dominate the cannibalization and service-cost burden, so the firm prefers the freemium strategy to premium price cuts as the means of expanding learning data. From a managerial perspective, this result highlights how consumer expectations about future capability improvement can strengthen premium pricing and make the freemium strategy more attractive over the long run. Figure 5 illustrates the resulting optimal strategy.

Figure 5. Optimal Freemium Decision in a Two-Stage Game



Notes. The figure depicts the firm's optimal strategy as a function of the service cost coefficient and outside option. Parameters are set to $q = 0.2$ and $\alpha = 0.8$, with qualitative patterns robust to alternative values.

4.2. Endogenous Base Model Capability

In the baseline model, we treat the free version's base model capability q as fixed and exogenous. In practice, however, an AI firm can choose which foundation model to deploy in the free version and can update it at different paces, making q a strategic design choice. We therefore extend the model by endogenizing q and assume the firm sets q before its freemium decision. Because a closed-form characterization is generally intractable, we rely on numerical optimization to derive the optimal q^*.

Observation 1. When the free version's base model capability q is endogenous and chosen before the firm makes the freemium decision,

(i) When α is relatively low, q^* is an interior solution.

(ii) When α is relatively high, q^* equals the upper boundary solution if s is low, but becomes an interior solution if s is high.

The firm's choice of q reflects the same three-way tradeoff highlighted at the beginning of Section 3.3. When α is small, the firm chooses an interior q^* because raising q mainly increases cannibalization and service costs, with only limited learning gains. When α is large, the stronger learning benefit creates a greater incentive to raise q, but this incentive is constrained when s is high because the additional usage induced by a higher q also increases service costs. Therefore, endogenizing q preserves the baseline logic while showing how the firm optimally designs the free version's capability to balance learning gains against cannibalization and service costs.

4.3. Nonlinear Adaptive Learning Increment Function

In the base model, θ^I is linear in ΔN. In practice, it is reasonable to expect diminishing marginal returns to additional training data, for example, due to constraints on compute, storage, and infrastructure scaling (Knight 2023). We now consider a concave specification, $\theta^I = \alpha \cdot \theta^B \cdot \sqrt{\Delta N}$, under which the marginal capability gain declines as the learning data set grows. Closed-form thresholds for comparing strategies are no longer tractable, so we rely on numerical analysis to characterize the optimal strategy and obtain the following observation.

Observation 2. When the marginal capability gain declines as the learning data set grows:

(1) Profit under the freemium strategy is higher than under the paid-only strategy when the value of the service cost coefficient s and the consumer's outside option c are relatively low.

(2) For the freemium strategy to be optimal, AI adaptive learning effectiveness α and the free version's base model capability q must fall within a narrower range than those of the base model.

Observation 2 shows that the baseline qualitative insight remains: Even with diminishing marginal returns to additional users, the freemium strategy can be optimal when both s and c are low. At the same time, diminishing marginal returns make the firm more reluctant to introduce a free version. Because each additional user may contribute less incremental capability improvement, the freemium strategy remains attractive only when cannibalization is limited (a lower q), and adaptive learning is sufficiently effective (a higher α). Relative to the baseline model, the region in which the freemium strategy dominates therefore becomes smaller.

Beyond robustness, this extension suggests a practical implication: As performance gains from additional data plateau (Knight 2023), mature firms with large data sets may benefit more from improving the effectiveness of adaptive learning (e.g., higher-quality data curation or refinement of fine-tuning methods) than from further expanding data set size.

4.4. Heterogeneous Preferences for Base and Adaptive Capabilities

In the base model, consumers have heterogeneous preferences for the base model capability θ^B, whereas preferences for the adaptive learning increment θ^I are homogeneous. In practice, however, consumers may differ substantially in how they value general-purpose capability from pretraining versus task- or domain-specific improvements from adaptive learning. To capture this feature, we extend the model to allow heterogeneity in preferences for both θ^B and θ^I. Let denote $0 < t_L < t_H < 1$ low and high valuations of θ^I, and let be λ the share of high-type consumers.

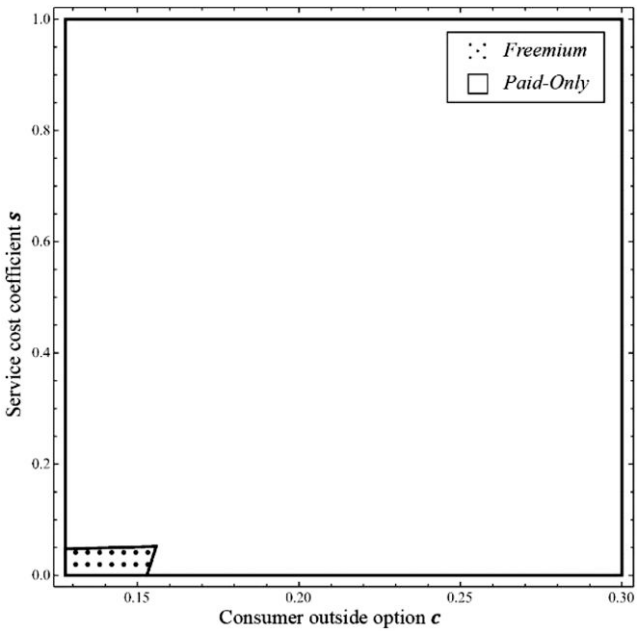
Proposition 6. *When consumers have heterogeneous preferences for both θ^B and θ^I, comparing profits under the freemium strategy (π^{V*}) and the paid-only strategy (π^{S*}):*

(i) *If $\frac{q\alpha t_H}{1+\alpha(1-\lambda)(t_H-t_L)} < c < \frac{(1-q)q}{2-q-\alpha\lambda t_H-\alpha(1-\lambda)t_L}$ and $\max\{0, s_5\} < s < s_6$, the firm prefers the freemium strategy, $\pi^{V*} > \pi^{S*}$;*

(ii) *Otherwise, the firm prefers the paid-only strategy, $\pi^{V*} < \pi^{S*}$.*

Here, s_5 and s_6 are defined in the Online Appendix. Compared with the base model, Proposition 6 shows that heterogeneity in preference for θ^I weakens the firm's incentive to offer a free version (the shaded area in Figure 6). The free version is valuable primarily because it generates interaction data that improves θ^I, thereby increasing the premium version's overall capability. When consumers value θ^I similarly, improvements can be broadly monetized through a higher willingness to pay. With heterogeneous preferences, a nontrivial share of consumers places relatively low value t_L on θ^I. As a result, increases in θ^I translate into smaller gains in willingness to pay for those users, reducing the marginal profit contribution of additional

Figure 6. Optimal AI Freemium Decision with Consumers' Heterogeneous Preferences for $\boldsymbol{\theta}^B$ and $\boldsymbol{\theta}^I$



Notes. The figure depicts the firm's optimal strategy as a function of the service cost coefficient and outside option. Parameters are set to $q = 0.3$, $\alpha = 0.8$, $\lambda = 0.6$, $t_L = 0.2$, and $t_H = 0.6$, with the qualitative patterns robust to alternative values.

interaction data and making the freemium strategy less attractive.

4.5. Outside Option for Both Versions

In the base model, we normalize the premium outside option to zero. In practice, consumers benchmark both versions against market alternatives: The free version competes with other no-cost or open-source tools, and the premium version competes with paid products offering similar capabilities. To capture this feature, we extend the model to allow outside options for both versions and examine how the premium outside option affects the firm's freemium strategy choice.

Let kc denote the outside option for the premium version, where $k > 0$ captures the premium's outside option relative to the free version. A higher (lower) k indicates the premium version faces a stronger (weaker) substitute than the free version. We also impose $0 < k < 1/c$ so that $kc < 1$ and interior solutions are preserved.

Proposition 7. *Consider the outside option for both versions, comparing profits under the freemium strategy* (π^{V*}) *and the paid-only strategy* (π^{S*}):

(i) *If* $q\alpha < c < c_3$ *and* $\max\{0, 2\alpha - 1 - ck\} < s < s_8$, *or*

(ii) *If* $c_3 < c < c_2$ and $s_7 < s < s_8$,

the firm prefers the freemium strategy, $\pi^{V*} > \pi^{S*}$; *otherwise,* $\pi^{V*} < \pi^{S*}$.

Proposition 7 shows that the main qualitative pattern persists when both versions face distinct outside options, as illustrated in Figure 7. Because introducing a premium outside option renders closed-form thresholds analytically intractable, we rely on numerical analysis to study how k influences the firm's strategy.

As k increases, the premium version faces stronger substitutes. Under a paid-only strategy, this tightens pricing constraints and limits the size of the paid user base even when service costs are relatively low, reducing both premium revenue and the learning data set. By contrast, under the freemium strategy, the free version can expand the user base and supplement the learning data set, partially offsetting cannibalization and helping preserve the premium version's value. As a result, the freemium strategy becomes profitable over a wider range of (c, s) as k increases.

4.6. Heterogeneous Service Cost Coefficient

In the base model, we assume a common service-cost coefficient s for both versions. In practice, however, service costs may differ across versions because premium users often submit more complex and resource-intensive queries. To capture this feature, we allow the free version to have a cost advantage relative to the premium version. Let $t \in (0, 1)$ denote the free version's service-cost parameter, where a lower t indicates that the free version is less costly to serve. We also assume that service costs are separable across versions, so this extension changes both the service-cost coefficients and the aggregation of service costs. Next, we rely on numerical analysis to characterize the conditions under which offering a free version is profitable.

Observation 3. The freemium strategy yields higher profits than the paid-only strategy when either

(i) The free version's service cost advantage is small, and both s and c are relatively low, or

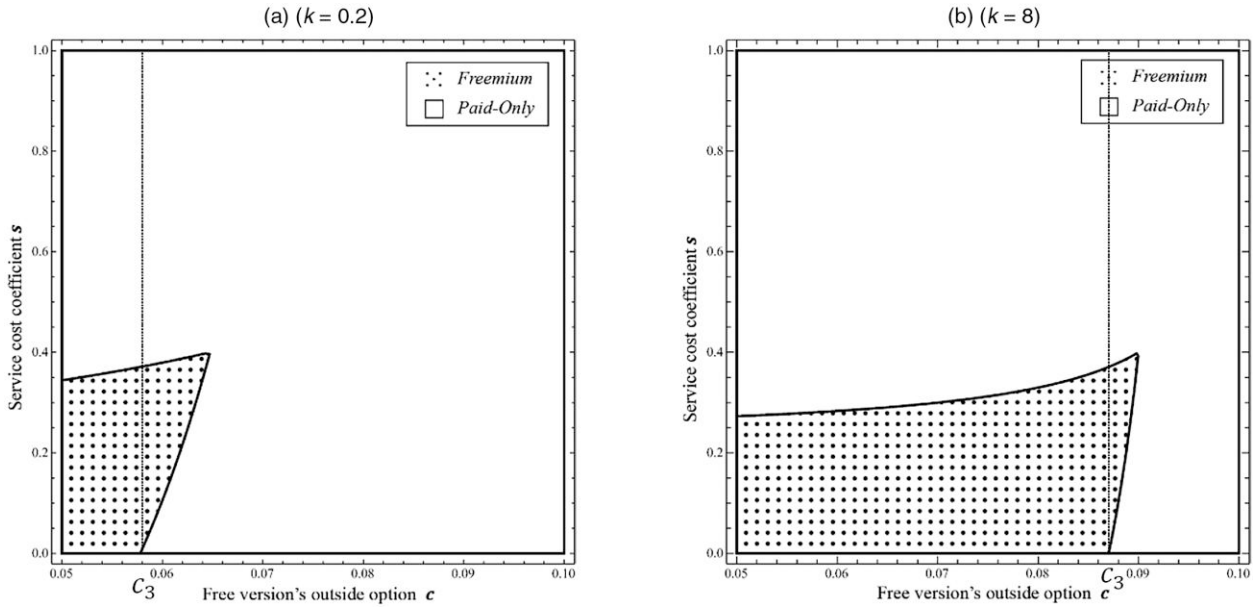
(ii) The free version's service cost advantage is large and c is low.

Observation 3 highlights how cost asymmetry shapes the firm's incentive to adopt the freemium strategy. When t is low, the free version is relatively inexpensive to serve, so the firm can attract additional free users at limited cost and benefit more from adaptive learning. As a result, the profitability of the freemium strategy becomes relatively insensitive to s (i.e., observation 3(ii)). In contrast, when t is high, serving free users becomes more costly, and the freemium strategy is profitable only when both s and c are low, aligning with the baseline tradeoff.

5. Conclusion

Generative AI is transforming how individuals and organizations create content and complete tasks. Yet the implications of learning-driven capability improvements for firms' monetization and freemium decisions

Figure 7. Optimal AI Freemium Decision Considering Outside Options for Both Versions



Notes. The figure depicts the firm's optimal strategy as a function of the service cost coefficient and the free-version's outside options. Parameters are set to $q = 0.1$ and $\alpha = 0.5$, with the qualitative patterns robust to alternative values.

remain underexplored. We address this gap by developing an analytical model that represents AI development as a two-stage process: In pretraining, AI acquires foundational capability from publicly available data; in fine-tuning, capability improves through AI adaptive learning driven by user interactions. We study how this learning process affects a firm's optimal strategy for offering free and premium versions and how it shapes profitability and consumer welfare.

Our analysis yields several insights. First, the freemium strategy is optimal only when AI adaptive learning effectiveness is in an intermediate range. If adaptive learning is too weak, free users generate limited capability improvements and mainly introduce cannibalization and service cost burdens. If adaptive learning is too strong, the free version can become too close in capability to the premium version, intensifying cannibalization. Second, when consumer outside options are moderate, firms may refrain from offering a free version even when service costs are low: The firm can instead cut the premium price to build a sufficiently large paid user base and generate learning data without cannibalization from a free version. As service costs rise to a moderate range, deep price cuts become less attractive, and a free version can become a more cost-effective way to expand the data set. Third, narrowing the base model capability gap between the free and premium versions can increase profits: A moderately stronger free version expands the user base, strengthens adaptive learning, and can raise the premium version's value enough to offset additional cannibalization and service cost burdens. Finally, on the consumer side, introducing a free version does not necessarily increase consumer surplus. When service costs are low, firms may leverage learning-driven capability gains to raise premium prices and appropriate a substantial share of the surplus, so total consumer surplus under the freemium strategy can fall below that under a paid-only strategy. Our extensions, a two-stage game, endogenous q, concave adaptive learning, heterogeneous preferences for θ^B and θ^I, heterogeneous service costs, and outside options for both versions, show that the qualitative insights are robust.

This paper contributes to two research streams. First, it contributes to the AI economics literature by providing an analytical framework that explicitly models learning-driven capability improvements through user interaction. Second, it contributes to the software versioning literature by showing how learning-driven interactions between versions, especially data generated by free users, can strategically enhance premium capability and change versioning incentives.

Our study also provides important insights for managers and policymakers in related industries. For AI firms aiming to optimize revenues through offering a free version, we recommend that they moderately narrow the AI capability gap to attract users and enrich AI adaptive learning data. Also, the firms should maintain balanced AI adaptive learning effectiveness to prevent excessive cannibalization. Additionally, we highlight that firms should carefully assess service costs and market conditions before introducing a free version. Specifically, with moderate consumer outside options, a free

version is advantageous only at moderate but not minimal service costs, where enriched user interactions effectively enhance AI capabilities via AI adaptive learning, offsetting cannibalization losses and boosting profitability. Finally, for policymakers, our finding that the freemium strategy can reduce consumer surplus highlights the need to monitor premium pricing and market power in large generative AI markets.

Our framework has several limitations that suggest promising directions for future research. First, we abstract from the broader AI value chain. A stylized AI ecosystem comprises three layers: infrastructure (compute and cloud services), model and platform (foundation models and access platforms), and applications (downstream products) (Susarla 2025). Our analysis focuses on monetization and freemium decisions at the model-and-platform layer and treats the other layers as exogenous. Future work could endogenize cross-layer value capture, for example, by examining settings in which firms offer models for free as a loss leader to stimulate downstream application adoption or to expand upstream infrastructure and cloud-service revenues. Second, within the model and platform layer, we restrict attention to subscription-based monetization. In our model, the free version matters primarily because it generates interaction data that raises premium willingness to pay. In practice, free usage may be monetized through additional channels (e.g., advertising, commerce/affiliate revenue or profits accruing in other layers). An important extension is to incorporate these alternative revenue functions and study how they reshape the optimal capability gap and the role of free access under adaptive learning. Third, our model is designed to capture capability evolution and strategic design choices rather than a repeated subscription process with churn and switching. A natural extension is a dynamic adoption-retention model in which users update beliefs after observing realized improvements and decide whether to upgrade or downgrade between versions over time. Such an extension would allow a richer treatment of churn, lock-in, and switching costs in AI subscription markets, complementing our focus on learning-driven versioning incentives.

Notwithstanding these limitations, this study provides an initial step toward understanding how AI adaptive learning interacts with product versioning and monetization incentives, and contributes to the growing literature on generative AI, digital platforms, and technology-enabled pricing strategies.

Acknowledgments

The authors thank the senior editor, the associate editor, and three anonymous reviewers for their valuable and constructive suggestions. L.Zhong acknowledges support from the China Postdoctoral ScienceFoundation [Grant KLH1019212]. Y. (R.) Tan acknowledges financial support from the CEIBS research funds [codes AG25TRD, BU26XIA, DP25TYL, and AU26BDA].

Endnotes

[1] For example, ChatGPT differentiates free and premium offerings through access to models with substantially different capability levels (OpenAI 2026a).

[2] Although model updates are continuous in practice, we adopt a stylized two-stage framework—pretraining and fine-tuning—to highlight capability development and user interaction. We refer to the user-interaction–driven refinement during fine-tuning as AI adaptive learning, a term used for expositional clarity.

[3] Although the pretraining and fine-tuning stages may overlap in practice, this abstraction is useful for modeling sequential strategic decisions regarding data acquisition, model improvement, and product differentiation.

[4] We use *versioning* as an umbrella term for offering multiple capability levels and analyze its freemium special case (free basic version and paid premium). The zero price of the free version is a modeling normalization; the tradeoff extends to low positive prices.

[5] A graphic designer may value a stronger base model for its creativity capabilities and image-handling functions, such as those of the GPT-5, whereas a student using the AI simply for language translation may have a lower appraisal of the same product; likewise, a privacy-sensitive user places a lower value on a given θ^B because they prefer outputs produced with fewer prompts and less cumulative disclosure.

[6] Continuing with the example of the graphic designer and the student, if the AI can better adapt to their specific tasks, regardless of the input content, they would value the learning increment θ^I more.

[7] Students seeking language translation services may have the following outside options: employing conventional translation software, translating by themselves, or switching from the current AI-based Service to another one developed by competing firms.

[8] Without AI adaptive learning, the firm does not offer a free version: a free version only cannibalizes premium demand and increases service costs. Thus, in our model, the incentive to offer a free version is driven by AI adaptive learning.

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. (2024) OpenAI. GPT-4 Technical Report. Retrieved July 24, https://cdn.openai.com/papers/gpt-4.pdf.

Berente N, Gu B, Recker J, Santhanam R (2021) Managing artificial intelligence. *MIS Quart.* 45(3):1433–1450.

Bhargava HK, Choudhary V (2008) Research note—When is versioning optimal for information goods? *Management Sci.* 54(5): 1029–1035.

Bloomberg (2024) Generative AI. Retrieved July 24, 2026, https://assets.bbhub.io/promo/sites/16/Bloomberg-Intelligence-NVDA-Gen-AIs-Disruptive-Race.pdf.

Brodsky S (2024) The hidden costs of AI: How generative models are reshaping corporate budgets. Retrieved July 24, 2026, https://www.ibm.com/think/insights/ai-economics-compute-cost.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, et al. (2020) Language models are few-shot learners. *Adv. Neural Inform. Processing Systems* 33:1877–1901.

Chellappa RK, Mehra A (2018) Cost drivers of versioning: Pricing and product line strategies for information goods. *Management Sci.* 64(5):2164–2180.

Chellappa RK, Shivendu S (2005) Managing piracy: Pricing and sampling strategies for digital experience goods in vertically segmented markets. *Inform. Systems Res.* 16(4):400–417.

Chen Y-J, Seshadri S (2007) Product development and pricing strategy for information goods under heterogeneous outside opportunities. *Inform. Systems Res.* 18(2):150–172.

Cheng HK, Liu Y (2012) Optimal software free trial strategy: The impact of network externalities and consumer uncertainty. *Inform. Systems Res.* 23(2):488–504.

Cheng HK, Li S, Liu Y (2015) Optimal software free trial strategy: Limited version, time-locked, or hybrid? *Production Oper. Management* 24(3):504–517.

Cui R, Li M, Zhang S (2022) AI and procurement. *Manufacturing Service Oper. Management* 24(2):691–706.

Feng H, Jiang Z, Liu D (2018) Quality, pricing, and release time: Optimal market entry strategy for new software-as-a-service vendors. *MIS Quart.* 42(1):333–353.

Feng J, Li X, Zhang XM (2019) Online product reviews-triggered dynamic pricing: Theory and evidence. *Inform. Systems Res.* 30(4):1107–1123.

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper. Management* 18(1):69–88.

Gao Y, Kumar S, Shen B, Tan Y (2025) The paradox of AI authentication: Better technology may not stop counterfeits. Preprint, submitted September 20, https://doi.org/10.2139/ssrn.4944305.

Google (2024a) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint, submitted March 8, https://arxiv.org/abs/2403.05530.

Google (2024b) An overview of the Gemini app. Retrieved July 24, 2026, https://gemini.google/overview/.

Google (2026a) Gemini apps privacy hub. Retrieved July 24, 2026, https://support.google.com/gemini/answer/13594961?visit_id=639202862954844278-1336800244&p=privacy_help&rd=1#pn_data_usage.

Google (2026b) Generative AI on Vertex AI training overview. Retrieved July 24, 2026, https://docs.cloud.google.com/vertex-ai/docs/training/genai-training-overview.

Gurkan H, de Véricourt F (2022) Contracting, pricing, and data collection under the AI flywheel effect. *Management Sci.* 68(12):8791–8808.

Hou T, Li M, Tan Y, Zhao H (2024) Physician adoption of AI assistant. *Manufacturing Service Oper. Management* 26(5):1639–1655.

Knight W (2023) OpenAI's CEO says the age of giant AI models is already over. *WIRED* (April 17), https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

Li M, Li T (2022) AI automation and retailer regret in supply chains. *Production Oper. Management* 31(1):83–97.

Loten A (2023) Rising data center costs linked to AI demands. *Wall Street Journal* (July 13), https://www.wsj.com/articles/rising-data-center-costs-linked-to-ai-demands-fc6adc0e.

Maslej N, Fattorini L, Perrault R, Gil Y, Parli V, Kariuki N, Capstick E, et al. (2025) Artificial intelligence index report 2025. Retrieved July 24, 2026, https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf.

OpenAI (2023) How should AI systems behave, and who should decide? Retrieved July 24, 2026, https://openai.com/blog/how-should-ai-systems-behave.

OpenAI (2026a) ChatGPT pricing. Retrieved July 24, 2026, https://chatgpt.com/pricing/?openaicom_referred=true.

OpenAI (2026b) Data usage for consumer services FAQ. Retrieved July 24, 2026, https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq.

OpenAI (2026c) Reinforcement fine-tuning. Retrieved July 24, 2026, https://developers.openai.com/api/docs/guides/reinforcement-fine-tuning.

Pei J, Wang R, Yan P, Tan Y (2025) Quality management in supply chain: Strategic implications and the paradox of AI inspection. *Decision Sci.* 56(5):509–526.

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, et al. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Machine Learn. Res.* 21(140):1–67.

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Shan G, Qiu L (2026) Examining the impact of generative AI on users' voluntary knowledge contribution: Evidence from a natural experiment on Stack overflow. *Inform. Systems Res.* 37(2):1021–1041.

Shivendu S, Zhang ZJ (2015) Versioning in the software industry: Heterogeneous disutility from underprovisioning of functionality. *Inform. Systems Res.* 26(4):731–753.

Susarla A (2025) Show me the money in AI: Innovation rails in the AI value chain. *Forbes* (October 8), https://www.forbes.com/sites/anjanasusarla/2025/10/08/show-me-the-money-in-ai-innovation-rails-in-the-ai-value-chain/.

Tan Y (2024) Implications of blockchain-powered marketplace of preowned virtual goods. *Production Oper. Management* 33(6):1393–1409.

Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, et al. (2023) Llama 2: Open foundation and fine-tuned chat models. Preprint, submitted July 18, https://arxiv.org/abs/2307.09288.

Vanian J, Leswing K (2023) ChatGPT and generative AI are booming, but the costs can be extraordinary. *CNBC* (March 13), https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html.

Wei XD, Nault BR (2014) Monopoly versioning of information goods when consumers have group tastes. *Production Oper. Management* 23(6):1067–1081.

Wu S-Y, Chen P-Y (2008) Versioning and piracy control for digital information goods. *Oper. Res.* 56(1):157–172.

Yinliang (Ricky) Tan is a professor of decision sciences and management information systems at the China Europe International Business School. He received his PhD in operations management and information systems from the University of Florida in 2015. His research focuses on artificial intelligence, digital platforms, the digital economy, and technology and operations management.

Ling Zhong is a postdoctoral researcher at Fudan University. She received her PhD in management science and engineering from Southwest Jiaotong University in 2024. Her research focuses on supply chain management, information systems, artificial intelligence, and the digital economy.

Michelle Y. Lu is an assistant professor of marketing at the China Europe International Business School. She received her PhD in marketing from Yale University. Her research focuses on artificial intelligence (AI) in the digital economy, AI and Big Data, content marketing, and applied game theory.

Jiajia Nie is a professor and vice dean of the School of Economics and Management at Southwest Jiaotong University. He received his PhD in management science and engineering from Chongqing University in 2009. His research focuses on logistics and supply chain management, decision optimization, service and operations management, and information systems.

Copyright of Information Systems Research (INFORMS) is the property of INFORMS: Institute for Operations Research & the Management Sciences and its content may not be copied or emailed to multiple sites without the copyright holder's express written permission. Additionally, content may not be used with any artificial intelligence tools or machine learning technologies. However, users may print, download, or email articles for individual use.