AI-delegated search: pricing, search quality, and business model implications*

Tat-How Teh† Hajime Shimao‡ Warut Khern-am-nuai§

July 18, 2026

Abstract

We study AI-delegated search: a consumer repeatedly queries an AI agent, which internally conducts costly sequential search and recommends options whose match values exceed a chosen threshold, incurring an API or compute cost for each evaluation. The interaction operates on two margins: the consumer decides whether to query repeatedly, while the provider sets the inner search-quality threshold. The provider's business model determines whether these margins are aligned. Under subscription pricing, consumers face a zero marginal price and over-query, which amplifies the provider's costs and induces the provider to distort search quality. The resulting distortion follows a double-crossing pattern: the threshold is too low when compute costs are very low or very high, and too high at intermediate cost levels. Under usage pricing, the per-query fee makes consumers account for the compute cost induced by their querying behavior, and the provider chooses the first-best threshold. Subscription pricing is more profitable when compute costs are low or competition is weak, whereas usage pricing dominates in the opposite cases. We further show how query caps, advertising, competition, and learning by AI agents can mitigate or amplify this alignment problem. Our analysis yields managerial and policy implications that differ from standard search-intermediary settings (e.g., search engines).

Keywords: AI agents, consumer search, business models, compute cost, competition

JEL Classification: D83, L86, L12, L13, L15

*Tat-How Teh gratefully acknowledges research funding from Singapore MOE-Tier 1 Seed Grant No. 023701-00001. We thank Ben Casner, Sanxi Li, Eeva Mauring, Markus Reisinger, Julian Wright, as well as other participants at the 24th ZEW Conference on ICT (Mannheim), the 15th Consumer Search and Switching Cost Workshop (Berlin), Shanghai University of Finance and Economics, Nanyang Business School, and Zhejiang University International Industrial Economics Conference for helpful comments. Errors are ours.

†Division of Economics, Nanyang Technological University. Email: tathow.teh@ntu.edu.sg.

‡Engineering Division, Penn State Great Valley. Email: hjs5825@psu.edu

§Desautels Faculty of Management, McGill University. Email: warut.khern-am-nuai@mcgill.ca

1 Introduction

Artificial Intelligence (AI) technologies, especially Generative AI (GenAI), have increasingly been integrated into many aspects of businesses (Bick et al., 2026). One of the most prevalent uses of GenAI is to assist consumers in searching for information (Immorlica et al., 2024). Unlike a traditional search engine, which returns a ranked list of search results for the user to evaluate, the AI agent[1] searches, screens the available options, and presents a recommendation.[2] Then, the user can inspect the recommendation and either accept it or ask the agent to try again. Behind this increasingly common interaction lies a set of economic and strategic questions that are central to AI markets. (1) Is this simply another reduction in consumer search costs, or does delegation to AI create a distinct search environment? (2) How much costly internal search should the AI provider perform before showing the recommendation? (3) How does the answer depend on whether the provider monetizes through a subscription, per-query usage fees, or advertising? (4) How do these business models shape the provider's incentives to design search quality, and how does competition among providers affect these choices?

We call this environment *AI-delegated search*: the consumer delegates a sequential evaluation task to an AI agent. This delegation couples two search margins. The consumer controls the *outer* margin, i.e., whether to inspect, accept, or query again. For each incoming query, the AI agent carries out the *inner* margin, i.e., performs searches internally and then shows a recommendation to the consumer. The provider controls this *inner* margin by choosing how selectively its AI agent performs internal searches, and bears the compute cost (e.g., API cost) of this internal search. Delegation therefore does more than lower the consumer's search cost; it makes the agent an active participant in the costly search process itself.

The presence of two search margins distinguishes AI-delegated search from standard search intermediation along three dimensions. First, AI-delegated search features a *double incidence of search costs*. The AI agent incurs a compute (API) cost to evaluate each option,[3] while the consumer incurs a human inspection cost to evaluate the recommended option. Every follow-up query therefore triggers costs on both sides of the relationship, in contrast to a search engine, whose cost of listing results does not significantly scale with how often a given consumer searches again. Second, the search process is interactive and shapes the provider's cost structure. Because a consumer can reject a recommendation and query again, prompting a fresh round of internal search, the provider's cost of acquiring information scales with the endogenous querying margin rather than being sunk or fixed. This feature makes the management of user querying behavior a central concern, and it contrasts with traditional intermediaries such as physicians or brokers.[4]

[1] Hereafter, we use the term *AI agent* for AI technologies that perform the search function. This terminology highlights that users delegate the search task to another entity rather than simply using AI as a tool. The underlying technology need not be fully autonomous in a strict technical sense. As such, AI-powered assistants and common GenAI tools also fall within our scope.

[2] In this paper, the term "recommendation" refers to search results that an AI agent shows to its users in response to the input query. Importantly, the focal technology in our study is search engine and not recommendation engine.

[3] We have in mind a "real-world" search whereby an AI agent searches on behalf of humans in an unstructured environment. This is different from an "on-platform search" where AI agents can, in effect, look for the globally optimal option at virtually zero cost (Hadfield and Koh, 2025).

[4] Traditional expert intermediaries incur costs of acquiring and processing information, but these costs are typically sunk or do not interact with an endogenous querying margin. The cost margin is therefore often abstracted from in models of search and advice provision, which instead emphasize search prominence, ordering rules, and information frictions (Armstrong et al., 2009; Inderst and Ottaviani, 2012).

Third, AI providers decide how to monetize users, which contrasts with the standard hiring arrangement whereby the entity carrying out the tasks (e.g., employees) is offered contracts by the delegator (e.g., a principal). The providers' choices of business model and pricing jointly drive user querying behavior and the search quality delivered, which are strategic levers absent from hiring relationships.

These features make the provider's business model (or monetization mode) a central strategic variable. In practice, consumer-facing plans have often relied on *subscription pricing* in which the consumer pays an upfront lump-sum fee and faces zero marginal price per query. Developer-facing APIs, by contrast, are typically priced on a per-query basis, which we call *usage pricing*. Meanwhile, some providers also monetize through sponsored placement or advertising. Table 1 summarizes these business models and provides stylized examples. These services are now offered at scale by providers such as Anthropic, OpenAI, and Google, yet there is still limited formal analysis of how AI-delegated search design interacts with the business model (Bergemann et al., 2026). The central question of this paper is therefore twofold: *given a business model, what drives the provider's choice of search-quality threshold; and, anticipating this threshold, which business model does the provider choose?*

Table 1: Examples of business models used by AI providers (as of June 2026)

Business model	**Description**	**Examples**
Subscription	Fixed monthly or annual fee; users face zero price per query but may face usage caps in some cases.	ChatGPT Plus; Google AI Pro (Gemini); Perplexity Pro; Claude Pro.[5]
Usage pricing	Per-query (or per-token) charge; users pay for each interaction.	GitHub Copilot; Google Vertex AI for enterprise search; API access for LLMs (e.g., those by OpenAI, Anthropic, and Google).[6]
Sponsored	Free to users; users see sponsored placement or embedded ads.	Microsoft Copilot; ChatGPT Free and Go tiers; Perplexity (withdrawn since 2026).[7]

We develop an economic model of AI-delegated search based on the canonical sequential search framework of Weitzman (1979). The model applies broadly to AI-assisted search for information on product matches (e.g., hotels or restaurants to visit) or solutions to tasks (e.g., suitable code snippets or draft documents). An AI agent sequentially screens the available options, observes a noisy signal of each option's match value, and recommends the first option whose signal exceeds a threshold chosen upfront by the provider. Hence, the agent's search quality is measured as the level of this inner-search threshold. Consumers inspect the search result at their own cost and decide whether to accept or query again. We begin with a monopoly provider that monetizes through either a flat subscription fee (*subscription pricing*) or a per-query usage fee (*usage pricing*). We then study query caps, advertising revenue, blind acceptance of the search result, and provider

[5]The Claude Pro subscription involves usage limits that scale with message and file length, with a five-hour reset cycle, in addition to aggregate weekly and monthly caps. Similar restrictions exist for ChatGPT but in a less restrictive manner. See Demirer et al. (2025) for evidence on capacity and pricing dynamics in the LLM market.

[6]GitHub Copilot historically used subscription pricing, but moved toward usage-based billing with GitHub AI Credits starting June 1, 2026; see GitHub (2026).

[7]Perplexity tested sponsored follow-up questions in 2024 but withdrew all advertising in early 2026, citing user-trust concerns; see Perplexity (2024) and Campaign (2026). Anthropic has explicitly declined to pursue advertising; see Boston Globe (2026). By contrast, Microsoft Copilot and ChatGPT (on its free and Go tiers, since 2026) continue to display ads; see AdVenture Media (2026).

competition as variations on this baseline.

The paper makes four contributions to research on business models and competitive strategy in AI-mediated markets. First, it identifies a new economic mechanism of consumer search in such markets: AI-delegated search couples user querying with the provider's costly inner-search efforts. In the efficient benchmark, the provider chooses a threshold that induces a high-acceptance outcome, i.e., the AI agent recommends an option that it expects the consumer to accept, to the best of its information. This result follows from the double incidence of search costs, whereby unnecessary follow-up queries waste both consumer inspection effort and provider compute cost. Conditional on inducing high acceptance, the efficient threshold then balances incremental informational gain against the marginal cost of internal search. This mechanism is distinct from standard search models in which the cost of acquiring information is sunk and does not scale with consumer usage.

Second, the paper shows that the provider's business model determines whether its search-quality design aligns with the efficient benchmark. Under usage pricing, the per-query fee makes consumers account for the marginal compute cost generated by their querying behavior, and the provider's threshold coincides with the efficient one. Under subscription pricing, consumers do not internalize compute cost and over-query; the provider responds by distorting the threshold to manage the inflated cost. The distortion has a double-crossing pattern, with the threshold too low when compute costs are very low or very high, but too high at intermediate cost levels. The reason is that the total cost is a function of the provider's cost per query and the number of user queries. Hence, the distortion pattern reflects two opposite ways for AI providers to control total cost: lowering the threshold to reduce cost per query versus raising it to reduce repeated queries. The latter suggests a counterintuitive managerial message, whereby the provider lowers total cost by providing better search quality.

As a point of comparison, in a traditional quality-investment model, a monopolist using a flat access fee chooses efficient quality and extracts surplus through the fixed fee. Meanwhile, in a typical search intermediation model, the intermediary affects ranking or prominence but does not bear an endogenous per-query compute cost. In our setting, neither logic applies because a flat fee separates the user's querying margin from the provider's compute-cost margin. Consequently, it not only distorts consumers' usage intensity but also causes distortions in search quality that can go in either direction.

Third, the paper derives business-model implications from this mechanism. In the monopoly benchmark, usage pricing is more profitable when consumers' direct-search alternative is sufficiently attractive or when compute cost is high, whereas subscription pricing dominates when compute cost is low and the provider is sufficiently differentiated. The comparison reflects a trade-off between usage pricing, which better aligns consumers' querying incentives, and subscription pricing, which extracts surplus through a lump-sum fee. Introducing query caps to subscription pricing, as has been done by some providers in Table 1, can mitigate over-querying, but they cannot fully replicate usage pricing because consumers' realized query demands are stochastic.

Fourth, the paper extends the analysis to competition between AI providers in a Hotelling duopoly setting. Competition preserves the monopoly characterization of the inner-search threshold but changes the business-model comparison. When horizontal differentiation between providers is sufficiently weak, usage pricing becomes the dominant strategy. This gives rise to a prisoner's dilemma where providers are jointly better off under subscription pricing. Yet, intense competi-

tion drives both to adopt usage pricing. In terms of strategic implications, the result shows that a provider's business model determines how aggressively it can compete for consumers. Usage pricing allows the provider to offer more value to consumers without sacrificing profit one for one, because lower per-query prices can stimulate additional usage. Hence, the viability of subscription pricing in consumer-facing AI markets depends on horizontal differentiation among providers. As AI agents become closer substitutes, the competitive pressure can trigger an industry-wide migration toward usage-based pricing.

We then extend the analysis in three directions. First, when the AI agent learns consumer preference along a consumer's query path (i.e., its signal noise declines with repeated queries), profits rise across all business models. Usage pricing remains aligned with the efficient threshold, whereas the subscription distortions remain. Second, allowing consumers to accept a recommendation without inspection (blind acceptance) partially weakens the double incidence of inspection costs. It can be efficiency-improving and sometimes mitigates the subscription distortion, by limiting the number of consumer queries much as a query cap does. Third, we let some of the AI agent's recommendations be sponsored and ask how disclosure of a recommendation's type affects the provider. Numerically, transparency creates a "routing benefit" whereby the provider assigns the organic and sponsored pools distinct roles along the query path, an option unavailable under opacity. This benefit is largest under subscription pricing, where advertising revenue also monetizes the excess querying that subscription induces.

Several implications emerge from our analysis. For managers, the key lesson here is that business models for AI-delegated search are not merely revenue instruments. They shape users' incentives to keep querying, providers' total costs, and providers' incentives to invest in internal search quality. These interactions imply that the optimal strategy to reduce total cost sometimes involves setting a high search-quality threshold, and that a query cap is not a perfect substitute for usage-based pricing. For policymakers, the analysis cautions that seemingly consumer-friendly flat pricing can create an over-querying externality that inflates compute costs and eventually results in distorted search quality delivered by AI. More generally, policies that regulate AI access, sponsored recommendations, or query limits should account for the provider's endogenous adjustment of pricing and search-design choices. We elaborate on these implications in Section 7.

The remainder of the paper proceeds as follows. Section 2 presents the model. Section 3 characterizes consumer querying, the efficient benchmark, and the provider's threshold choices under usage and subscription pricing. Section 4 compares the business models. Section 5 analyzes competition between AI providers. Section 6 develops three extensions: AI learning along query path, blind acceptance, and sponsored-option transparency and opacity. Section 7 discusses managerial and policy implications, and then concludes.

1.1 Literature

Search and intermediation. This paper contributes to the literature on consumer search and intermediation rooted in the seminal work of Weitzman (1979). Most clearly related are those considering settings with search engines or experts who provide recommendations (e.g., financial intermediaries and physicians).[8] A series of analytical studies model the search engine as a tech-

[8]Another strand of this literature studies information design and recommendations by multi-sided retail platforms that facilitate transactions between buyers and sellers and monetize from the transactions, such as Amazon and eBay

nology that reduces consumers' effective search cost or dictates their search sequence, and examine issues such as advertising pricing (Eliaz and Spiegler, 2011), placement auctions (Athey and Ellison, 2011; Chen and Zhang, 2018), and biases that arise when the engine has commercial interests in the products it lists (De Cornière and Taylor, 2014, 2019). On the expert-recommendation front, the focus has been on how commission-based compensation distorts expert recommendations and equilibrium market prices (Armstrong and Zhou, 2011; Inderst and Ottaviani, 2012; Teh and Wright, 2022). In a slightly different vein, Janssen and Williams (2024) consider an expert influencer who acquires information upfront by sampling available products and recommends the best among them to consumers, examining the resulting market outcomes. A common theme across these studies is that the intermediary derives its compensation from the "supply side" (advertisers and firms selling products and services).

Our setting departs in two dimensions. First, rather than producing a one-shot ranking or recommendation, the AI agent runs repeated inner-search loops that respond instantly to user queries. The provider's compute cost therefore scales with the endogenous number of queries, giving rise to marginal-cost considerations that are absent from this literature. Second, in our setting the AI provider monetizes primarily from the consumer side through subscription and usage pricing, and the interaction between these business models and user querying behavior is at the core of our analysis.

Delegated search. A related literature studies delegated search as a principal–agent problem, where a principal induces a better-informed or effort-constrained agent to conduct sequential search on her behalf (Armstrong and Vickers, 2010; Erat and Krishnan, 2012; Lewis, 2012; Ulbricht, 2016; Zorc et al., 2025). Their focus is on optimal principal–agent contracting (e.g., delegation sets, bonus schemes, and stopping rules) to discipline an agent whose search effort or information is non-contractible. Our focus is instead the opposite direction of contracting. Specifically, the AI provider offers the contract in the form of a business monetization model, whereas the consumer, typically the "principal" in this literature, is passive. Because the provider both designs and monetizes the search, the analysis must confront issues that do not arise in traditional delegation models, such as users' excessive querying, usage caps, and advertising.

Economics of autonomous AI. A growing body of work treats AI not as a transformative technology per se, but as an entity that acts on behalf of its users (Hadfield and Koh, 2025). A strand of this literature considers AI agents used by product sellers, i.e., algorithms that set prices for firms (Calvano et al., 2020; Fish et al., 2024). We belong instead to the strand examining AI agents used by consumers and developers in solving a variety of workflows and tasks, especially the recent theoretical contributions by Mahmood (2024) and Bergemann et al. (2026) that examine monetization by providers of such AI agents.[9] Mahmood (2024) formulates AI providers as allocators of compute tokens for solving a variety of tasks, and examines competition among providers but restricts attention to linear token pricing. Bergemann et al. (2026) show that token-budget (usage caps) menus, two-part tariffs, and committed-spend contracts are equivalent implementations of the same optimal non-linear mechanism, because the buyer's type pins down total demand

(Hagiu and Wright, 2020; Zhong, 2023; Zou and Zhou, 2025; Karle et al., 2026; Ke et al., forthcoming).

[9] Demirer et al. (2025) provide a comprehensive survey on the current pricing practices and market shares of the leading AI providers.

at the ex-ante contracting stage.[10]

Relative to these contributions, our paper studies AI providers and agents in a sequential search environment that captures the interactive way consumers use AI agents. In this setting, query demand is not fixed at the contracting stage; it is an interim random variable generated by the user's accept-or-query-again decisions. This distinction breaks the equivalence between usage caps and two-part tariffs considered by Bergemann et al. (2026). The reason is that a cap constrains realized search only after the stochastic sequence unfolds, whereas a per-query price changes the user's marginal querying decision at every step.

2 Baseline model

The environment consists of a unit mass of unit-demand "consumers", a unit mass of "options" (which could be products or services that consumers potentially want, answers, or variants of solutions to consumers' tasks), and an AI provider selling access to AI agents that assist consumers in search.

Consumer search without AI. A representative consumer's match value for each option i is denoted by $v_i \in [\underline{v}, \bar{v}]$, which is drawn i.i.d. across options. Let F be the cumulative distribution function (CDF) of v_i, with full support on $[\underline{v}, \bar{v}]$ and a log-concave density f. The latter implies an increasing hazard rate and a decreasing mean residual life (Bagnoli and Bergstrom, 2005):

$$\frac{\partial}{\partial v}\left[\frac{f(v)}{1-F(v)}\right] \geq 0 \quad \text{and} \quad \frac{\partial}{\partial v}\left[\frac{\int_v^{\bar{v}}[1-F(v_i)]\,dv_i}{1-F(v)}\right] \leq 0. \tag{1}$$

Meanwhile, consumers' value from doing nothing is normalized to zero. Consumers are initially uninformed about their match values with potential options and can learn them only through costly sequential search. Each search incurs a direct inspection cost $s_0 > 0$, which we interpret as the cost of scrutinizing the details of the option and exercising human judgment.

In the absence of the AI agents, consumers inspect a randomly drawn option at each step of search and then decide whether to accept the option or continue searching. This captures traditional search through a search engine such as Google or Bing, or, more generally, manually exploring solutions to tasks at hand. Consumers have free and perfect recall.

Search through AI agents. If a consumer signs up with the AI provider, then at each step of search she can send a query to the provided AI agent, receive a recommendation, and then *inspect the recommended option* at an inspection cost $s > 0$. After inspecting the recommended option, the consumer decides whether to accept it or continue querying the AI agent. As is standard in the search literature, the consumer has to inspect the option before accepting it.

To generate recommendations, the AI agent conducts an internal random sequential search,

[10] In a slightly different vein, another line of work examines generative AI overviews that synthesize informational content across multiple sources, rather than the autonomous AI agents searching for products, services, or solutions that we consider here. Ng and Wessel (2026) study Google's strategic deployment of such AI Overviews as a form of platform self-preferencing, while Liang et al. (2025) and Kaiser et al. (2025) provide empirical evidence on how users interact with generative AI relative to traditional search engines in search-related tasks.

sequentially drawing and inspecting options i to observe a partially informative signal $\hat{v}_i$ that follows a "truth-or-noise" structure introduced by Lewis and Sappington (1994) and further developed by Johnson and Myatt (2006). That is, with probability $1-\mu$ the signal $\hat{v}_i$ is informative and equals the true value v_i, and with probability μ the signal is an independent uninformative draw from the same distribution F.[11] Therefore, the parameter $\mu \in (0,1)$ captures the noisiness in the AI agent's signal, or more generally, the misalignment between its preferences and the consumer's.

For each option internally inspected, the AI agent incurs a compute cost $c > 0$. The AI agent then recommends the first encountered option with a signal $\hat{v}_i$ above a chosen inner-search threshold t, which is an upfront design choice announced by the AI provider.[12] Apart from its tractability, the threshold-based rule is also natural on practical grounds. In modern search and recommendation systems, retrieval involves using machine learning models to score the relevance between a query and a large set of available options, surfacing only those options clearing a relevance threshold (Guo et al., 2022). The same structure describes AI agents handling open-ended tasks, such as identifying an appropriate hotel or restaurant for a trip. A parallel logic applies when AI agents produce solutions to workflow tasks: an AI agent commonly iterates on a candidate solution (generating, evaluating, and refining it in an evaluator-optimizer loop) until the output clears a completeness threshold.[13] Our inner-search threshold t plays the role of these relevance and completeness thresholds.[14]

We initially consider two basic business models. In the *subscription* model, the provider charges a fixed participation fee $T \geq 0$ for signing up and nothing per query. In the *usage* model, consumers sign up for free but pay a per-query fee p for each recommendation. In our model, the per-query fee p is *outcome-equivalent* to usage-based pricing that is based on AI workload (or token consumption), as practiced by API pricing in Table 1. To see this formally, suppose that for each recommendation the consumer instead pays ρ per step of inner search performed to deliver it. Each inner search step stops with probability $1-F(t)$. Hence, at the point of querying, the expected payment per query is $\frac{\rho}{1-F(t)}$. Since the provider chooses the threshold t, this is equivalent to the provider choosing the per-query payment $p = \frac{\rho}{1-F(t)}$ directly.

Timing.

1) The AI provider sets the subscription fee T (subscription model) or the per-query fee p (usage model), and the inner-search threshold t of its AI agent.
2) Consumers observe the provider's choices and decide whether to sign up with the AI provider. Then, consumers search sequentially.

[11] This signal structure implies the unconditional distribution of the signal $\hat{v}_i$ is also F, which simplifies the notation but does not otherwise affect the qualitative features of the results below. See also Hagiu and Wright (2020) and Guo (2023) for applications of the truth-or-noise signal framework to recommendation settings.

[12] We treat t as an observable commitment: it is a single design parameter of the deployed agent, applied uniformly across users and queries. Because it governs realized quality for all users at once, ex-post degradation is detectable in the aggregate via, e.g., third-party benchmarks and user reports of "quality drops" (e.g., VentureBeat, 2026). Therefore, it is disciplined by the provider's concern for future participation and reputation.

[13] See, e.g., Anthropic, "Building Effective Agents," December 2024, https://www.anthropic.com/engineering/building-effective-agents; and Google Cloud, "Choose a Design Pattern for Your Agentic AI System," May 2025 https://docs.cloud.google.com/architecture/choose-design-pattern-agentic-ai-system.

[14] The threshold-based search rule is equivalent to stopping when the conditional expected value satisfies $\mathbb{E}[v_i|\hat{v}_i] \geq \mathbb{E}[v_i|t]$. The formulation also appeared in Zhong (2023) with finite n products. However, Zhong (2023) assumes that the intermediary's signals are perfect to gain tractability in analyzing product sellers' pricing incentives, whereas we allow for noisy signals but abstract away seller pricing.

We solve for a subgame perfect Nash equilibrium. We assume $\max\{s_0, s\} < \mathbb{E}[v_i] - \underline{v}$, where $\mathbb{E}[\cdot]$ is the expectation operator, ensuring that direct search is a viable outside option and ruling out a zero-search outcome. Throughout, all stand-alone qualifiers such as "positive," "increasing," and "log-concave" are understood in the weak sense.

Modeling discussions. Three clarifications about the baseline are in order.

First, the baseline assumes ex-ante identical consumers for expositional clarity, but our main results are unchanged when consumers have heterogeneous outside options (which, in our model, is equivalent to heterogeneous s_0). We formally analyze this in the duopoly analysis of Section 5 via a Hotelling framework that nests local monopolies with downward-sloping consumer participation.

Second, our baseline is deliberately minimal to isolate the two search margins that distinguish AI-delegated search. In doing so it abstracts from richer features of how consumers interact with AI agents. We revisit two such features later. First, the AI agent may learn from a user along the query sequence, reducing its signal noise μ the more the user queries (Section 6.1). Second, a consumer may blindly accept a recommendation without verifying it, skipping the inspection cost s (Section 6.2).

Third, the baseline considers pure usage and subscription pricing for clarity. We later consider two-part tariffs and subscription with usage caps in Section 4, and study advertising-based revenue in Section 6.3.

3 Equilibrium analysis

3.1 Consumer decisions

Search directly. Suppose the consumer does not sign up with the AI provider. In this case, the consumer searches directly by herself. By the standard results in sequential search (Weitzman, 1979), the corresponding search reservation value $v_0 \in (\underline{v}, \bar{v})$ is defined by

$$\int_{v_0}^{\bar{v}} (1 - F(v_i))\, dv_i = s_0.$$

The consumer optimally stops at the first option whose value exceeds v_0, and so the expected number of steps of search is $\frac{1}{1-F(v_0)}$. Then, a standard identity under the reservation rule implies

$$\mathbb{E}[v_i | v_i \geq v_0] - \frac{s_0}{1 - F(v_0)} = v_0.$$

The reservation value v_0 serves as the effective outside option for consumers in this model. That is, it is the expected utility if a consumer chooses not to sign up with the provider.

Search with AI agent. If a consumer has signed up with the AI provider, then at each step of search (querying), her AI agent recommends an option whose signal meets the inner-search threshold $\hat{v}_i \geq t$. By Bayes' rule, the consumer's posterior over the match value of the recommended

option is

$$\begin{aligned} G(v|t) &= \Pr(v_i \leq v | \hat{v}_i \geq t) \\ &= \mu F(v) + (1-\mu)\frac{F(v)-F(t)}{1-F(t)} \cdot \mathbf{1}_{\{v \geq t\}}, \quad v \in [\underline{v}, \bar{v}]. \end{aligned}$$

Here, $\mathbf{1}_{\{.\}}$ is the indicator function, equal to 1 when its argument holds and 0 otherwise.

To proceed, define the consumer's incremental gain function as

$$\Phi(v|t) \equiv \int_v^{\bar{v}} (1 - G(v_i|t))\, dv_i \quad \text{for } v \in [\underline{v}, \bar{v}].$$

This function captures consumers' querying incentives, as the value $\Phi(v|t)$ is each consumer's incremental gain from querying through AI when the current best option at hand is v. Note that $\Phi(v|t)$ is decreasing in v and, by first-order stochastic dominance, increasing in t.

Given per-query fee p (with $p = 0$ capturing the subscription model), the search reservation value $r(t,p)$ is defined by equating the incremental gain and the incremental cost of searching:

$$\Phi(r(t,p)|t) = s + p. \tag{2}$$

The consumer stops querying after seeing an option value exceeding $r(t,p)$.[15] Again, by standard identities, $r(t,p)$ equals the expected utility a consumer obtains if she signs up with the provider and queries the AI agent at every step of search. Summarizing:

Lemma 1. *In every step of search, the consumer prefers searching through the AI (as compared to searching directly) if and only if $r(t,p) \geq v_0$, or equivalently,*

$$p \leq \Phi(v_0|t) - s.$$

Moreover, $r(t,p)$ is continuous, strictly decreasing in p, and strictly increasing in t.

Thus, as long as the per-query fee is not too large, searching through an AI agent dominates searching directly, because searching sequentially through AI generates a better distribution of options than a direct sequential search.

A key object in our analysis is the expected number of queries by a consumer, defined as

$$\alpha(t,p) \equiv \frac{1}{1 - G(r(t,p)|t)},$$

i.e., the reciprocal of the acceptance probability $1 - G(r(t,p)|t)$. Notably, $\alpha(t,p)$ is kinked because the posterior distribution G is piecewise, reflecting that consumers exhibit two qualitatively different acceptance patterns depending on whether the threshold is high enough to satisfy $t \geq r(t,p)$.

Lemma 2. *Consider a consumer who searches through an AI agent. There exists a boundary threshold $\hat{t}$ such that the following holds:*[16]

[15] The definition of reservation value in equation (2) implicitly assumes $s + p < \Phi(\underline{v}|t)$, i.e., p is not too high, which indeed holds in equilibrium. For completeness, whenever $s + p \geq \Phi(\underline{v}|t)$, then the consumer stops at the first recommendation, and so we set $r(t,p) = \underline{v}$ without loss of generality. Therefore, $r(t,p) \in [\underline{v}, \bar{v})$.

[16] We label the boundary case $t = \hat{t}$ as a high-acceptance regime without loss of generality, since it is a kink point of both $r(t,p)$ and $\alpha(t,p)$. Labeling the boundary case as a low-acceptance regime does not change our results.

(i) *(Low-acceptance) if* $t < \hat{t}$*, then* $t < r(t,p)$*, with* $\alpha(t,p)$ *strictly decreasing in* t *on* $[\underline{v}, \hat{t})$.

(ii) *(High-acceptance) if* $t \geq \hat{t}$*, then* $t \geq r(t,p)$*, with* $\alpha(t,p)$ *strictly increasing in* t *on* $[\hat{t}, \bar{v}]$.

Lemma 2 says that the expected number of queries is U-shaped in t, with the kink point $\hat{t}$ defined by $\hat{t} = r(\hat{t}, p)$. Intuitively, raising t has two effects: (i) it improves the posterior distribution of recommended options (G becomes higher in first-order stochastic dominance), which implies higher acceptance probability; and (ii) it raises the consumer's reservation value r, which lowers acceptance probability as the consumer becomes pickier.

In the low-acceptance regime ($t < \hat{t}$), the reservation value is a smooth function and a direct calculation shows that the expected number of queries can be expressed as a mean residual life:

$$\alpha(t,p) = \frac{1}{s+p}\left(\frac{\int_r^{\bar{v}}(1-F(v_i))\,dv_i}{1-F(r)}\right),$$

which is decreasing in r under log-concavity. So, a higher t, which forces r upward (Lemma 1), implies fewer queries. In this case, the improvement in the posterior distribution dominates.

The logic changes in the high-acceptance regime ($t \geq \hat{t}$). In this regime, the inner-search threshold t is above the reservation value r, meaning that consumers reject options *only if* the AI agent's signal is incorrect, and so

$$\alpha(t,p) = \frac{1}{1-\mu F(r)}$$

which is increasing in r and hence t. In this regime, consumers' acceptance is driven by realizations where the AI agent's signal is incorrect, the conditional distribution of which is not responsive to t. Therefore, an increase in t just makes the consumer pickier, prolonging the querying sequence. In the limiting case of $\mu \to 0$, the high-acceptance regime entails $\alpha(t,p) \to 1$; i.e., the consumer accepts the first recommended option.

Before proceeding, it is useful to summarize the key notation from this preliminary analysis:

Table 2: Key notation

Symbol	Meaning
s_0	Per-inspection cost of consumer when searching directly
v_0	Consumer reservation value when searching directly
s	Per-inspection (search) cost of consumer when querying through AI
c	Per-evaluation compute cost incurred by the AI provider
t	Stopping threshold of the inner-search loop by the provided AI agent
$r(t,p)$	Consumer reservation value when using AI at per-query fee p and threshold t
$\alpha(t,p)$	Expected number of queries by a consumer, $\alpha(t,p) = [1 - G(r(t,p)\|t)]^{-1}$
$\Phi(v\|t)$	Incremental gain from querying with best-so-far v
$\hat{t}$	Boundary threshold: $r(\hat{t},p) = \hat{t}$; separates low- and high-acceptance regimes

3.2 First-best efficiency benchmark

We define the first-best as the outcome that maximizes total welfare. In the presence of the AI provider, total welfare is

$$\mathrm{TW}(t,p) = r(t,p) + \left(p - \frac{c}{1-F(t)}\right)\alpha(t,p),$$

consisting of the consumer's expected utility plus the provider's revenue minus the total compute cost.

The first-best requires consumers' search decisions to fully internalize the true cost of queries, so the efficient per-query fee equals the expected marginal compute cost: $p = \frac{c}{1-F(t)}$. Welfare maximization then reduces to maximizing $r(t, \frac{c}{1-F(t)})$ over t. In what follows, denote the first-best solution as t_{eff} and $r_{\text{eff}} \equiv r(t_{\text{eff}}, \frac{c}{1-F(t_{\text{eff}})})$.

Proposition 1 (Efficiency benchmark)**.** *The following hold:*

(i) *If $\mu c < (1-\mu)s$, then the first-best threshold t_{eff} is the unique solution to*

$$(1-\mu)\int_{t_{\text{eff}}}^{\bar{v}} (1-F(v_i))\,dv_i = c, \tag{3}$$

and it induces a high-acceptance regime where $t_{\text{eff}} > r_{\text{eff}}$.

(ii) *If $\mu c \geq (1-\mu)s$, then the first-best induces a low-acceptance regime where $t_{\text{eff}} = \underline{v} \leq r_{\text{eff}}$.*

Moreover, AI-delegated search improves upon direct search if and only if $r_{\text{eff}} \geq v_0$.

The efficiency benchmark exhibits a "bang-bang" feature. When the signal is not too noisy (case i), the efficient outcome is a high-acceptance regime. That is, the AI makes a recommendation that it believes consumers will accept (i.e., the option is rejected only if the AI agent's signal is incorrect). The reason is that delegated search inherently involves a "double incidence" of inspection costs. The AI agent incurs at least c to generate a recommendation and the consumer incurs s to inspect it, so every query costs at least $c+s$. To reduce this duplication, the outcome is such that consumers are expected to query infrequently. Observe that the threshold t_{eff} is decreasing in the noise parameter μ. That is, a noisier signal leads the AI provider to set a lower bar, "checking in" with the consumer more frequently throughout the query sequence.[17]

Meanwhile, when the signal becomes sufficiently noisy (case ii), the first-best jumps discontinuously from (3) to $t_{\text{eff}} = \underline{v}$, with minimal inner search by the AI agent. Intuitively, when the AI agent cannot reliably infer the consumer's match value, its internal search adds cost without proportional informational benefit. As such, the AI agent simply checks in with the consumer at every step of search. The outcome then reduces to standard sequential search at effective cost $s+c$, and this is less efficient than direct search if and only if $s + c > s_0$.

To focus on the economically interesting case, we impose the following assumption for the remainder of the paper, unless otherwise stated.

[17] It is worth emphasizing that the double-incidence logic, rather than the condition $c < s$, drives the efficient threshold in case (i). To see this, suppose $c > s$ and the AI agent is still used. A low threshold, say $t = \underline{v}$, leads to multiple instances of the combined cost $c+s$ before acceptance; a high threshold $t \geq r$ is more efficient as it incurs multiple instances of c but only a few instances of s before acceptance.

Assumption A. *The AI agent's compute cost c satisfies $\mu c < (1-\mu)s$ and $r_{\text{eff}} \geq v_0$.*

Assumption A requires that AI-delegated search is sufficiently efficient and focuses on case (i), where the first-best threshold is characterized by an interior solution (3). As noted above, when $\mu c \geq (1-\mu)s$, AI-delegated search is equivalent to a quantitative change in the search cost. The provider optimally chooses $t = \underline{v}$ across all business models, trivializing the design problem.

The following example based on the uniform distribution illustrates Proposition 1 and Assumption A. We will sometimes refer to this leading example to illustrate and sharpen some of our results. Online Appendix A collects all the details of this example.

Example 1. *Suppose F is uniform on $[0,1]$ and $s \leq s_0 < \frac{1}{2}$. In the direct search benchmark, $v_0 = 1 - \sqrt{2s_0}$. Under Assumption A, case (i) of Proposition 1 applies and the first-best outcomes are:*

$$t_{\text{eff}} = 1 - \sqrt{\frac{2c}{1-\mu}}, \qquad r_{\text{eff}} = \frac{1}{\mu}\left(1 - \sqrt{(1-\mu)^2 + 2\mu\sqrt{2c(1-\mu)} + 2\mu s}\right).$$

When $\mu \to 0$, $t_{\text{eff}} = 1 - \sqrt{2c}$ and $r_{\text{eff}} = 1 - \sqrt{2c} - s$, with $r_{\text{eff}} > v_0$ if and only if

$$c < \tfrac{1}{2}(\sqrt{2s_0} - s)^2.$$

3.3 AI provider's pricing and threshold choices

We now analyze the AI provider's decisions in each business model. We start by noting the following profit benchmark.

Remark 1. *The highest achievable profit is $r_{\text{eff}} - v_0$, i.e., first-best total welfare minus the consumer outside option, and it can be achieved via a two-part tariff, namely a subscription fee $T = r_{\text{eff}} - v_0$ together with per-query fee $p = \frac{c}{1-F(t_{\text{eff}})}$.*

In practice, AI providers adopt distinct pure pricing models for distinct market segments: usage pricing for API/developer access, flat subscription for consumer-facing plans (Table 1). We take these observed forms as the object of study and analyze each in isolation to characterize how the business model shapes the AI provider's recommendation design and market outcomes. Clean two-part tariffs with a marginal price from the first query are relatively uncommon. Where usage-based elements do enter consumer plans (e.g., Claude Max plans), they take the form of included allowances with overage top-ups, i.e., caps rather than a standing marginal price, which we analyze in Section 4.2. As such, we use Remark 1 only as an efficiency benchmark. Our main question is which of the two observed pure forms a provider prefers and how that business model shapes its search-quality design.[18]

[18] Several reasons might explain the absence of a usage-based component in consumer-facing plans. First, a standing marginal charge from the first query requires transparent metering of the AI agent's internal workload, which is costly to communicate to non-technical consumers (as opposed to developers). Second, consumers exhibit a well-documented aversion to marginal charges relative to lump-sum fees (Prelec and Loewenstein, 1998; Yun and Suk, 2022), which further pushes consumer-facing monetization toward flat access.

Usage pricing model

Under this business model, if a consumer participates then the expected number of queries is $\alpha(t,p) = \frac{1}{1-G(r(t,p)|t)}$. Therefore, the provider chooses (t,p) to maximize profit

$$\left(p - \frac{c}{1-F(t)}\right) \cdot \alpha(t,p),$$

subject to participation constraint $r(t,p) \geq v_0$ (Lemma 1). Given that $p \mapsto r(t,p)$ is a bijection as shown in (2), we can conveniently reformulate the provider's problem as choosing a threshold t and a target reservation value (or utility level) $u = r(t,p)$, implemented through p. So, the profit function becomes

$$\Pi_{\text{usage}}(t,u) = \left(\Phi(u|t) - s - \frac{c}{1-F(t)}\right) \cdot \frac{1}{1-G(u|t)}, \quad \text{subject to } u \geq v_0. \tag{4}$$

Solving the profit maximization delivers the equilibrium outcome:

Proposition 2 (Usage pricing)**.** *The equilibrium inner-search threshold is $t_{\text{usage}} = t_{\text{eff}}$, inducing reservation value $u_{\text{usage}} \in [v_0, r_{\text{eff}}]$ and a high-acceptance outcome. Moreover, t_{usage} is strictly decreasing in c and μ.*

Proposition 2 says that usage pricing aligns the AI provider's incentives in search design. The key reasoning is that the provider can always use per-query fee p to implement its target level of consumers' reservation value (in this case, $r(t,p) = u$), regardless of its choice of t. Then, with reservation value fixed via choosing p, raising t has two effects that exactly mirror the planner's trade-off: it improves the quality of each recommendation but increases the internal search cost per recommendation. Because the consumer's querying behavior is already disciplined by the fee, the AI provider faces no additional distortion and chooses the efficient threshold $t_{\text{usage}} = t_{\text{eff}}$.

Meanwhile, the participation constraint in (4) may be binding ($u_{\text{usage}} = v_0$) or slack ($u_{\text{usage}} > v_0$) at optimum. Intuitively, it reflects a classic price-vs-demand trade-off in pricing. Offering a higher reservation value (hence utility) to consumers requires the provider to charge a lower per-query price p, but it also encourages more queries, and hence potentially more profit.

Finally, both monotonicities have a natural interpretation. A higher compute cost c raises the per-query cost $\frac{c}{1-F(t)}$, so the AI provider lowers the inner-search threshold to economize. Likewise, a higher noise μ degrades signal quality, making frequent check-ins with the consumer more valuable, which again pushes the threshold down.

Subscription pricing model

This model is equivalent to $p = 0$. The AI provider optimally sets the lump-sum fee T equal to the consumer's maximum willingness to pay, $r(t,0) - v_0$, i.e., the increment in expected utility compared to searching directly. Once signed up, the expected number of queries is $\alpha(t,0) = \frac{1}{1-G(r(t,0)|t)}$, and the AI provider's profit is

$$\Pi_{\text{subs}}(t) = r(t,0) - v_0 - \frac{c}{1-F(t)} \cdot \alpha(t,0). \tag{5}$$

The equilibrium outcome is:

Proposition 3 (Subscription pricing)**.** *Suppose F has a weakly concave mean residual life.*[19] *There exist cutoffs $0 < \underline{c} < \bar{c}$ such that the equilibrium inner-search threshold t_{subs} satisfies:*

(i) *If $c < \underline{c}$, then $t_{\text{subs}} < t_{\text{eff}}$, inducing a high-acceptance outcome.*

(ii) *If $c \in (\underline{c}, \bar{c})$, then $t_{\text{subs}} > t_{\text{eff}}$, inducing either a high- or a low-acceptance outcome.*

(iii) *If $c > \bar{c}$, then $t_{\text{subs}} < t_{\text{eff}}$, inducing a low-acceptance outcome.*

Moreover, t_{subs} is decreasing in c.

Proposition 3 is the main threshold-distortion result. Under subscription pricing, the provider extracts surplus through a lump-sum fee but does not price the marginal query. Consumers therefore do not internalize the compute cost of each additional query. At the efficient threshold $t = t_{\text{eff}}$, consumers are over-querying, so the provider must use the inner-search threshold t to control the cost consequences of over-querying.

The resulting distortion is not monotone and reflects two distinct cost-control routes. To see this, recall the total cost

$$\frac{c}{1 - F(t)} \cdot \alpha(t, 0)$$

is a product of per-query cost $\frac{c}{1-F(t)}$ and total expected queries α. A lower threshold reduces the expected number of inner searches before each recommendation, thus lowering the per-query cost. However, at the same time it also reduces recommendation quality, which can either decrease or increase the number of queries by Lemma 2, depending on whether a high-acceptance or a low-acceptance regime prevails. Therefore, the provider can lower total cost by either: (i) lowering t to make each query cheaper; or (ii) adjusting t (possibly upward) to induce early acceptance.

In more detail, when c is *small*, both $\frac{c}{1-F(t)}$ and α are decreasing in t, so the provider opts for lowering per-query cost by choosing $t \leq t_{\text{eff}}$ while retaining high acceptance. At *intermediate* c, lowering t further would cross into the low-acceptance regime, where rejection rates and number of queries jump. The provider therefore sets a high threshold $t > t_{\text{eff}}$ to preserve high acceptance and avoid a high α. Finally, at *high* c, sustaining a high threshold becomes prohibitively expensive, so the AI provider sets $t < t_{\text{eff}}$ despite the resulting high α, because the per-query cost savings now dominate.

The intermediate region reflects a managerial message. In a standard quality-investment setting, lowering total cost typically calls for a lower quality. That logic fails here because quality also changes the number of queries α. The subscription provider may therefore choose a threshold above the efficient level because higher quality acts as a cost-control device that reduces costly repeat queries, rather than because higher quality is intrinsically overvalued.

Figure 1 illustrates Propositions 2 and 3 by plotting equilibrium thresholds under usage and subscription pricing for the stated parameters. The first panel shows that usage pricing coincides with the first-best threshold, whereas subscription pricing crosses the efficient threshold twice.

[19] Weak concavity of mean residual life is satisfied by the class of Generalized Pareto distributions with non-positive shape parameter, which nests the uniform and exponential distributions. Without it, the bidirectional distortion still survives, with $t_{\text{subs}} < t_{\text{eff}}$ at $c < \underline{c}$; $t_{\text{subs}} > t_{\text{eff}}$ at $c \in (\underline{c}, \bar{c})$; and $t_{\text{subs}} < t_{\text{eff}}$ for sufficiently large c. Hence, the technical condition is used only to guarantee that, when $c > \bar{c}$, search quality crosses the efficient level at most once. Extensive numerical simulations based on the truncated normal (which does not satisfy the condition) have not produced any such additional crossings.

The second panel shows that both business models generally yield lower profit than the first-best two-part tariff benchmark (Remark 1). We compare the two business models and their variations in the next section.

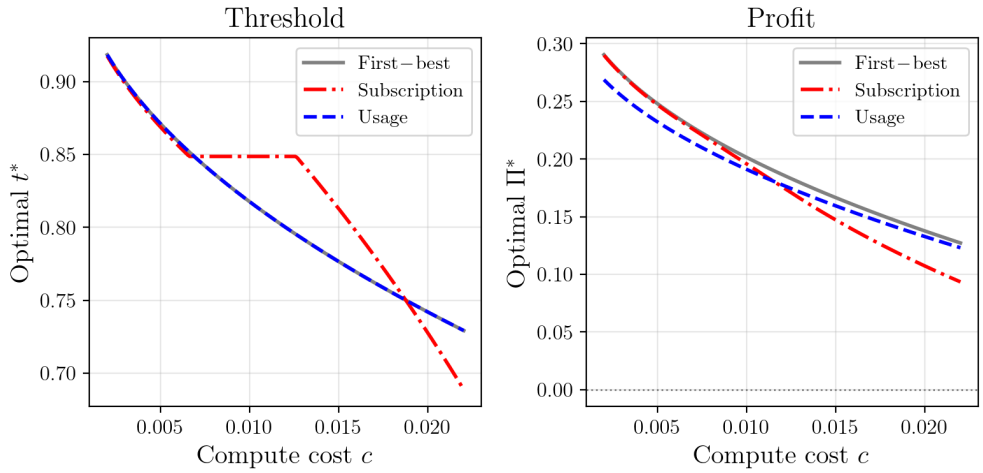

Figure 1: Equilibrium thresholds and profits under usage and subscription pricing. The parameters are $F \sim U[0,1]$, $s = 0.05$, $s_0 = 0.1$, and $\mu = 0.4$.

4 Comparison of business models

This section compares business models using the threshold characterizations above. We first contrast usage pricing and subscription pricing in terms of induced search quality and provider profit. We show that introducing query caps to the baseline subscription model does not affect the main insights from the comparison.

4.1 Usage pricing versus subscription pricing

Search threshold comparison

We now ask which business model induces a higher inner-search threshold, and hence better search quality, for a given cost environment. The following result is an immediate consequence of Propositions 2 and 3:

Corollary 1. *Under the condition stated in Proposition 3, the following comparisons hold:*

(i) *If* $c < \underline{c}$ *or* $c > \bar{c}$, *then* $t_{\text{usage}} > t_{\text{subs}}$;

(ii) *If* $c \in (\underline{c}, \bar{c})$, *then* $t_{\text{usage}} < t_{\text{subs}}$.

The corollary yields testable predictions. The comparison of search quality across the two models crucially depends on the parameter c, which admits two natural empirical interpretations.

First, holding computational costs fixed in the short run, c varies across *use-case domains*, since different tasks require the AI to consume different amounts of computation per internal evaluation. Some applications involve relatively low per-evaluation cost, for example, product comparison or routine information retrieval, where each candidate can be screened with a short bout of AI reasoning. Others demand substantially more computation per evaluation, for example,

professional-services matching or complex workflows that require multi-step reasoning. Under this cross-sectional reading, the corollary predicts that subscription pricing delivers superior search quality in moderate-cost domains, whereas usage pricing dominates in domains at both the low-cost and high-cost ends.

Second, holding the use case fixed, c can be viewed as declining over time as AI inference technology improves. Under this dynamic reading, the corollary suggests that an AI provider currently operating in the intermediate-cost region (where subscription delivers higher search quality) will eventually transition into the low-cost region as c falls below $\underline{c}$, at which point usage pricing delivers strictly higher search quality.

AI provider profit

We now compare equilibrium profit across the two business models. An important caveat is that, in practice, factors outside the model may also influence this comparison. Most notably, subscription pricing avoids the psychological cost of repeated per-use payments and offers consumers a simplicity benefit that usage pricing does not.[20] The comparison below should therefore be understood as isolating the role of the cost and signal-quality environment, holding such behavioral considerations constant across business models.

Proposition 4 (Profit comparison)**.** *The following comparisons hold:*

(i) *There exists a cutoff* $v_{\text{profit}} \in [\underline{v}, r_{\text{eff}})$ *on the outside-option value such that*

$$\Pi^*_{\text{usage}} > \Pi^*_{\text{subs}} \Leftrightarrow v_0 > v_{\text{profit}}.$$

(ii) *If compute cost* $c > 0$ *is sufficiently small (large), then* $\Pi^*_{\text{usage}} < (>)\Pi^*_{\text{subs}}$.[21]

The profit comparison between usage and subscription pricing reflects a business-model trade-off between incentive alignment and surplus extraction. Usage pricing has an advantage in aligning consumer incentives as it disciplines consumer querying and supports the efficient inner-search threshold. Its drawback is that extracting profit through a per-query markup also changes the consumer's outer querying decision. Subscription pricing has the opposite profile. A lump-sum fee is a powerful surplus-extraction instrument because it does not directly tax marginal queries, but the absence of a marginal query price leads to over-querying and inflates the provider's cost.

This trade-off gives a simple managerial lesson. Usage pricing is favored when incentive alignment is valuable or when surplus-extraction is less important. This happens when compute costs (c) are high or when competitive outside options (v_0) limit the provider's ability to extract surplus. Conversely, subscription pricing is favored when the cost of misalignment is relatively low.

Proposition 4(i) formalizes the outside-option channel. In a more competitive environment (higher v_0), the efficiency advantage of usage pricing dominates. To see why, we apply the envelope

[20] A large literature in behavioral economics and consumer behavior documents that consumers exhibit "pain of paying" effects that are stronger under per-use charges than under lump-sum fees. See, e.g., Prelec and Loewenstein (1998) and, more recently, Quispe-Torreblanca et al. (2019).

[21] In the proof, we show that a sufficiently large c here means c is such that $r_{\text{eff}} \to v_0$, and so it does not violate Assumption A. In Online Appendix A, we obtain a cutoff if-and-only-if characterization by focusing on Example 1. Formally, if μ is small enough and $F \sim U[0,1]$, then there exists a cutoff $c_{\text{profit}} > 0$ such that $\Pi^*_{\text{usage}} < \Pi^*_{\text{subs}} \Leftrightarrow c < c_{\text{profit}}$. Figure 2 illustrates this cutoff characterization for μ.

theorem to (4) and (5) and simplify:

$$\frac{d}{dv_0}\Pi^*_{\text{subs}} = -1$$

whereas $\frac{d}{dv_0}\Pi^*_{\text{usage}} = 0$ if the participation constraint is non-binding, and otherwise

$$\frac{d}{dv_0}\Pi^*_{\text{usage}} = -1 + \left(\Phi(v_0|t) - s - \frac{c}{1-F(t)}\right) \cdot \frac{\partial}{\partial v_0}\left(\frac{1}{1-G(v_0|t)}\right) > -1. \tag{6}$$

Intuitively, as v_0 rises, the provider must concede surplus to consumers by lowering its fees. Under subscription pricing, this translates into a one-for-one reduction in profit because the fee is a lump-sum charge. Under usage pricing, however, the lower per-query fee encourages additional querying, so part of the lost margin is offset by higher query volume. The intuition is clearest at the boundary case $v_0 = r_{\text{eff}}$. Under usage pricing, the provider can still break even by setting $p = \frac{c}{1-F(t_{\text{eff}})}$. Subscription pricing, however, becomes unprofitable because the provider's profit equals $\max_{t\in[\underline{v},\bar{v}]} \text{TW}(t,0) - v_0 < r_{\text{eff}} - v_0 = 0$, where the inequality reflects the inherent over-querying problem under subscription pricing.

Proposition 4(ii) formalizes the cost channel. When the compute cost c is small, the surplus-extraction advantage of subscription pricing dominates its efficiency disadvantage. The inefficiency of subscription pricing arises because consumers face $p = 0$ and do not internalize compute cost. When c is small, each excess query is cheap, so the total surplus loss from over-querying is modest. The provider can then benefit from the clean surplus extraction of a lump-sum fee. Conversely, when c is large, the excessive queries are costly, making subscription pricing less profitable.

Figure 2 illustrates Proposition 4 by indicating parameter combinations where $\Pi^*_{\text{usage}} > \Pi^*_{\text{subs}}$ (blue) and $\Pi^*_{\text{usage}} < \Pi^*_{\text{subs}}$ (white). Consistent with the proposition, usage pricing dominates when the outside option is strong or compute cost is high, whereas subscription pricing dominates when the provider is differentiated and compute cost is low.

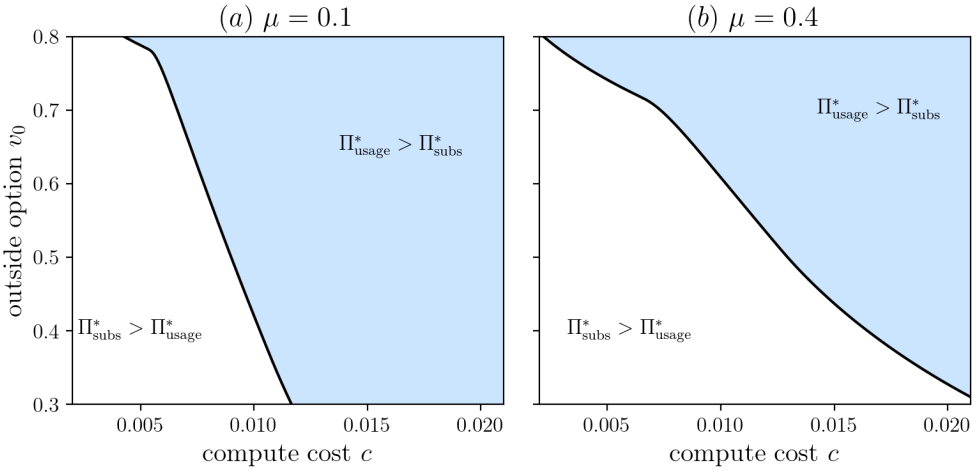

Figure 2: Comparison of AI provider profit across business models. The parameters are $F \sim U[0,1]$, $s = 0.05$, and $\mu = 0.1$ and $\mu = 0.4$.

We close this subsection with two further remarks. First, optimal pricing by providers means that in equilibrium consumer surplus is exactly v_0 under subscription pricing but generally weakly higher than v_0 under usage pricing, as the provider sometimes leaves surplus to consumers to prolong the query sequence. Consumer surplus is therefore weakly higher under usage pricing.

Consequently, whenever $\Pi^*_{\text{usage}} \geq \Pi^*_{\text{subs}}$, usage pricing yields higher total welfare.

Second, signal noise is necessary for subscription pricing to dominate. As $\mu \to 0$, $\Pi^*_{\text{usage}} \to r_{\text{eff}} - v_0 \geq \Pi^*_{\text{subs}}$. One might then conjecture that higher μ is generally more favorable for subscription pricing. However, numerical simulations show the relationship is ambiguous. Higher μ can either expand or contract the region $(\underline{c}, \bar{c})$ of Corollary 1, and both cutoffs in Proposition 4 vary non-monotonically in μ. Thus, no general conclusion can be drawn about how the threshold and profit comparisons change with the level of signal noise.

4.2 Subscription with cap

A natural managerial response to over-querying under subscription pricing is to impose a usage cap, which preserves the consumer-facing simplicity of a flat price while giving the provider a second instrument for controlling the number of user queries. In what follows, we show that caps can only partially replicate the incentive alignment generated by usage pricing.

In this subscription-with-cap model, the AI provider additionally chooses a *query cap* $n \in \{1, 2, \ldots\}$. The consumer pays a subscription fee T upfront and may submit at most n queries at no additional charge. Beyond n queries the service terminates and the consumer takes the best AI recommendation seen so far or reverts to direct search. With a slight abuse of notation, we allow $n = \infty$, indicating that the provider imposes no query cap and the model reduces to the baseline subscription model.[22]

The query cap simply transforms the consumer's querying problem into a finite-option search environment. By standard results for sequential search with i.i.d. draws and perfect recall (Weitzman, 1979), the consumer's stopping rule remains a cutoff rule with reservation value $r(t, 0)$, regardless of how many queries remain. However, the cutoff $r(t, 0)$ no longer equals the expected consumer utility. In what follows, we write $r \equiv r(t, 0)$ for brevity.

Consumer WTP for subscription-with-cap. To explicitly state the expected consumer utility from signing up, let $q = G(r(t, 0)|t)$ denote the single-draw rejection probability. In the *cap-binding event* (probability q^n), all n draws fall below r. Let $V_{(n)} = \max\{v_1, \ldots, v_n\}$; the consumer then takes $\max\{V_{(n)}, v_0\}$. So

$$\mathbb{E}[\max\{V_{(n)}, v_0\} \,|\, V_{(n)} < r] = v_0 + \int_{v_0}^{r} \left(1 - \frac{G(v_i|t)^n}{q^n}\right) dv_i.$$

In the *non-binding event* (probability $1 - q^n$), the consumer accepts the first draw exceeding r. The expected accepted value (excluding inspection costs) is

$$\mathbb{E}[v_i | v_i \geq r] = r + \frac{s}{1 - q};$$

[22] In Online Appendix B, we consider instead a cap N on the total amount of inner searches that the AI agent performs for the consumer, capturing a cap on total token consumption seen in practice. We show that it is equivalent to imposing a stochastic query cap n that follows a binomial distribution. Under the assumption that the inner-search sequence within any given query never stops prematurely (i.e., the AI agent always delivers recommendations with $\hat{v}_i \geq t$, which ensures stationarity in consumer querying), the insights from Proposition 5 continue to hold.

Meanwhile, a consumer queries m times if the first $m-1$ draws are all below r, and so the probability of using at least m queries is q^{m-1}. The total expected number of queries is

$$\alpha = \sum_{m=1}^{n} q^{m-1} = \frac{1-q^n}{1-q}\,.$$

Combining both events (weighted by the corresponding probabilities) and subtracting the expected search cost αs yields the consumer's maximum willingness to pay for a subscription-with-cap plan (t, n):

$$r \;-\; v_0 \;-\; \int_{v_0}^{r} G(v_i|t)^n \, dv_i. \tag{7}$$

Equation (7) shows that the willingness to pay and expected number of queries are both increasing in the cap n, which is intuitive. If $n \to \infty$, the willingness to pay converges to $r - v_0$ and $\alpha \to \frac{1}{1-q}$, consistent with the baseline subscription model.

AI provider choices. By equation (7), we can write the profit under subscription-with-cap as a perturbation of pure subscription profit:

$$\Pi_{\text{cap}}(t,n) = \Pi_{\text{subs}}(t) \;\; - \underbrace{\int_{v_0}^{r} G(v_i|t)^n \, dv_i}_{\text{WTP loss}} + \underbrace{\frac{q^n\, c}{(1-q)(1-F(t))}}_{\text{cost saving}}.$$

Let t_{cap}, n_{cap}, and Π^*_{cap} denote the optimal choices and profit.

Proposition 5 (Subscription with a query cap)**.** *Under subscription pricing with a query cap $n \in \{1, 2, \ldots\}$, the equilibrium satisfies the following properties:*

(i) *If $c \to 0$, then $n_{\text{cap}} \to \infty$ and $t_{\text{cap}} \to t_{\text{subs}}$. If c is large enough, then $n_{\text{cap}} = 1$ and $t_{\text{cap}} = t_{\text{eff}}$.*

(ii) *If v_0 is large enough, $\Pi^*_{\text{cap}} < \Pi^*_{\text{usage}}$; if c is small enough, then $\Pi^*_{\text{cap}} > \Pi^*_{\text{usage}}$.*

Proposition 5(i) reflects the following trade-off. A tighter cap reduces the expected number of queries and hence compute expenditure, but it also lowers the consumer's willingness to pay. The cost-saving consideration strengthens with c while the WTP loss is independent of it. Therefore, when $c \to 0$ the WTP consideration dominates and the provider optimally implements the baseline subscription outcome ($n_{\text{cap}} \to \infty$ and $t_{\text{cap}} \to t_{\text{subs}}$). Likewise, when c is large, the cost-saving consideration dominates and the provider imposes a tight cap ($n_{\text{cap}} = 1$), and the corresponding first-order condition for the threshold coincides with that of the first-best so that $t_{\text{cap}} = t_{\text{eff}}$.

Proposition 5(i) thus shows that the query cap acts as a partial substitute for usage pricing. At the two extremes, the provider implements either the unconstrained subscription threshold or the efficient one. Whether t_{cap} moves monotonically between these endpoints as c varies requires verifying joint comparative statics that are analytically difficult; we confirm the monotone pattern numerically.

Proposition 5(ii) resembles the profit comparison in Proposition 4. It highlights that the cap does not fully substitute for per-query pricing in addressing consumers' failure to internalize compute costs. Subscription-with-cap can never fully implement the first-best outcome, because the stochastic nature of consumer search means the realized number of queries the consumer would optimally choose can differ from any deterministic cap.

5 Competing AI providers

The monopoly analysis assumes consumers are ex-ante homogeneous so that in equilibrium the AI provider attracts all consumers. In this section, we consider competing providers to obtain further insights on how competition shapes equilibrium threshold choices and business model adoptions.

Suppose two AI providers $k \in \{1, 2\}$ are ex-ante symmetric, each deciding its business model $\omega_k \in \{subs, usage\}$. The providers are located at endpoints $\ell_1 = 0$ and $\ell_2 = 1$ of a unit line. A unit mass of consumers is distributed on $[0, 1]$ according to a CDF $H(x)$ with density $h(x)$ (symmetric around $x = 1/2$), and transportation cost $\tau > 0$ per unit distance. We assume h is log-concave, which ensures quasi-concavity of each provider's profit.

The timing parallels the monopoly baseline. In Stage 0, both providers simultaneously choose their business models; in Stage 1, both providers simultaneously choose their fees and inner-search thresholds; in Stage 2, consumers choose which provider to sign up with and engage in sequential search. We continue to impose Assumption A.

A consumer at location x who signs up with provider k obtains expected utility

$$u_k - \tau\,|x - \ell_k|, \quad \text{where } u_k = \begin{cases} r(t_k,\, p_k) & \text{if } \omega_k = usage \\ r(t_k,\, 0) - T_k & \text{if } \omega_k = subs \end{cases} \tag{8}$$

Then, the indifferent consumer is located at $\frac{1}{2} + \frac{u_1 - u_2}{2\tau}$, giving demands $D_1 = H(\frac{1}{2} + \frac{u_1 - u_2}{2\tau})$ and $D_2 = 1 - D_1$. We maintain the full-coverage assumption throughout. That is, v_0 is small enough that every consumer strictly prefers signing up with one of the two providers over the outside option of direct search.

Invariance of threshold choices

Each provider k's profit is expressed as

$$\begin{aligned} \Pi_k(t_k,\, p_k) &= \left[p_k - \frac{c}{1 - F(t_k)}\right] \cdot \frac{1}{1 - G(u_k|t_k)} \cdot D_k && \text{if } \omega_k = usage. \\ \Pi_k(t_k,\, T_k) &= \left[T_k - \frac{c}{1 - F(t_k)} \cdot \frac{1}{1 - G(r(t_k, 0)|t_k)}\right] \cdot D_k && \text{if } \omega_k = subs, \end{aligned}$$

The expressions reveal an important *separability* property. The provider's threshold choice can be decoupled from the competitive dimension. Under usage pricing, we can reparameterize a provider k's choice variables from $(t_k,\, p_k)$ to $(t_k,\, u_k)$ given the one-to-one relation in (8). Under this change of variables, demand D_k depends on u_k but not on t_k, because competition operates entirely through the utility offer, and so $\Pi_k(t_k,\, u_k) = \pi_{\text{usage}}(t_k, u_k) \cdot D_k$, where the per-consumer profit is

$$\pi_{\text{usage}}(t_k, u_k) = [\Phi(u_k|t_k) - s - \frac{c}{1 - F(t_k)}] \cdot \frac{1}{1 - G(u_k|t_k)},$$

which resembles the monopoly profit function (4) after replacing u_k with v_0. Likewise, under subscription pricing, we can reparameterize from $(t_k,\, T_k)$ to $(t_k,\, u_k)$ and so $\Pi_k(t_k,\, u_k) =$

$\pi_{\text{subs}}(t_k, u_k) \cdot D_k$, where the per-consumer profit is

$$\pi_{\text{subs}}(t_k, u_k) = r(t_k, 0) - \frac{c}{1 - F(t_k)} \cdot \frac{1}{1 - G(r(t_k, 0)|t_k)} - u_k,$$

which again resembles the monopoly profit function (5). Therefore, the monopoly benchmark applies. Each provider k's choice of threshold is independent of its choice of utility offer u_k.[23]

Corollary 2. *In any equilibrium of the Hotelling game, the inner-search threshold under each business model is invariant to the competitive environment. That is, if a provider has $\omega_k = usage$, then it optimally chooses t_{usage} as stated in Proposition 2; if a provider has $\omega_k = subs$, then it optimally chooses t_{subs} as stated in Proposition 3.*

The separability property arises in our setting because improvement in search quality involves a *marginal* cost. The threshold cost is per-consumer as it arises from the AI agent's per-query search effort, which makes it separable from the competitive dimension. Notably, this separability does not arise in settings where search quality improvements are fixed-cost investments (e.g., investing in a better search database or improving ranking algorithms). In those settings, the investment incentives would depend on the market share of the provider and hence the competitive environment.

To state our next result, we denote $\Pi^*(\omega_k; \omega_{-k})$ as provider k's equilibrium profit when it adopts business model ω_k and the opponent adopts ω_{-k}.

Proposition 6 (Business-model choice under competition)**.** *Suppose $\tau > 0$ is sufficiently small, then the following ranking holds*

$$\Pi^*(usage; subs) \geq \Pi^*(subs; subs) > \Pi^*(usage; usage) \geq \Pi^*(subs; usage).$$

Consequently, adopting the usage-pricing model is a weakly dominant strategy.

Proposition 6 carries two messages. First, providers behave more competitively when adopting usage pricing, because of how the pricing instrument affects competitive markups. The reasoning is the same as (6). Under subscription pricing, offering one additional unit of u_k to consumers entails a one-unit reduction in profit; under usage pricing, it entails a *less-than-one-unit* reduction in profit due to changes in users' querying behavior.

Second, the profit ranking indicates a prisoner's dilemma. Providers earn a strictly higher profit if both of them adopt subscription pricing, but it is individually rational for each to adopt usage pricing. Intuitively, when τ is sufficiently small and the market is highly competitive, in equilibrium all consumers sign up with the provider who can offer a higher consumer utility. By the same reasoning as in Proposition 4, a usage-pricing provider can always offer a higher utility due to its efficiency advantage. Therefore, in any asymmetric configuration with $\omega_k \neq \omega_{-k}$, the usage-pricing provider wins the entire market, and so $\Pi^*(\text{usage}; \omega_{-k}) \geq \Pi^*(\text{subs}; \omega_{-k})$ regardless of the opponent's business model.

[23] When full coverage fails, each provider operates as a local monopolist serving consumers within distance $\frac{u_k - v_0}{\tau}$ of its endpoint. In this regime, the threshold invariance of Corollary 2 continues to hold by the same separability argument, because demand in the local-monopoly regime also depends only on the utility instrument and not on t_k.

In the opposite case of sufficiently large τ, providers are local monopolists and the monopoly comparison in Proposition 4 applies. Figure 3 plots the equilibrium outcome for intermediate ranges of τ, confirming the same theme as the monopoly profit comparison. Competition (small τ) strengthens the appeal of usage pricing, whereas weaker competition (differentiation, large τ) strengthens the appeal of lump-sum subscription pricing.

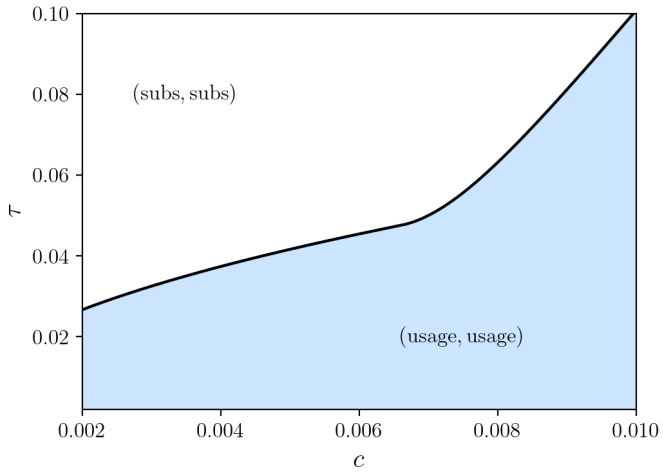

Figure 3: Equilibrium business models by competing providers. The blue (white) region indicates parameter combinations where mutual adoption of usage pricing (subscription pricing) is the equilibrium. The parameters are $F \sim U[0,1]$, $s = 0.05$, $s_0 = 0.1$, and $\mu = 0.4$.

6 Extensions

The baseline holds the AI agent's signal noise fixed, has the consumer inspect every recommendation before accepting it, and abstracts from advertising. This section develops three extensions: (i) the agent learns along the query sequence, so its signal noise falls with repeated use (Section 6.1); (ii) the consumer may accept a recommendation without inspecting it (Section 6.2); and (iii) some recommendations are sponsored, and the provider chooses whether to disclose which ones (Section 6.3). In each case we present the main insights here and relegate the formal details to Online Appendices C.1–C.3.

6.1 Learning and decreasing AI signal noise

So far the AI agent's signal noise μ has been fixed. In practice, as a consumer interacts repeatedly, the agent accumulates context about her preferences and its recommendations sharpen. We capture this by letting the signal noise decline along the consumer's query sequence, and ask how such learning reshapes the equilibrium.

Consumer search. Index the consumer's steps by $m = 1, 2, 3, \ldots$, where the beginning of step m is the decision point just before her m-th query. The signal noise of the recommendation returned by the m-th query is μ_m, where $\{\mu_m\}_{m \geq 1}$ is a decreasing sequence in $[0, 1)$ with $\mu_1 = \mu$, and

$$\mu_1 \geq \mu_2 \geq \mu_3 \geq \cdots \geq 0.$$

This captures the agent learning from repeated interaction. As it accumulates context about the consumer's preferences, its signal becomes weakly more informative query after query. The constant schedule $\mu_m \equiv \mu$ recovers the baseline. Every baseline object inherits the step index: the consumer's posterior over the m-th recommendation is

$$G_m(v|t) = \mu_m F(v) + (1-\mu_m)\frac{F(v)-F(t)}{1-F(t)}\,\mathbf{1}_{\{v\geq t\}},$$

with incremental gain $\Phi_m(v|t) \equiv \int_v^{\bar{v}}(1-G_m(v_i|t))\,dv_i$. The provider still commits to a single inner-search threshold t.

Because the environment is no longer stationary, the consumer's reservation value varies across steps. Let r_m denote her reservation value at the beginning of step m, and $r_\infty \equiv \lim_m r_m$.

Lemma 3. *Fix the inner-search threshold t and per-query fee p. The consumer's optimal search is a step-dependent reservation rule $\{r_m\}_{m\geq 1}$ defined by*

$$r_m = r_{m+1} + \Phi_m(r_{m+1}|t) - (s+p), \qquad m = 1,2,3,\ldots. \tag{9}$$

At the beginning of step m she accepts her realized best-so-far value v_i if and only if $v_i \geq r_m$, and otherwise issues the m-th query. Moreover, the sequence r_m is increasing in step count m.

The sign-up value is r_1, the expected utility of querying through the AI agent, so participation against direct search again requires $r_1 \geq v_0$. Once a consumer starts querying, she will keep querying until eventually accepting a recommended offer. The search reservation value is increasing in step count, i.e., $r_m \leq r_{m+1}$. Intuitively, as the agent learns, each successive recommendation is drawn from a better distribution, so the consumer optimally becomes more demanding the longer she has searched. In the degenerate case of constant noise, $r_m = r_{m+1}$ so that (9) recovers the baseline identity (2).

Equilibrium under learning. Online Appendix C.1 formally sets up the provider's problem, and details its numerical solution. For the numerical analysis we adopt the parametric schedule

$$\mu_m = \frac{\mu}{m^\eta}, \qquad \eta \geq 0,$$

where η is a *learning rate* indexing how fast the agent's recommendations sharpen with repeated use ($\eta = 0$ is the constant-noise baseline $\mu_m \equiv \mu$). We solve the provider's problem for the uniform example $F = U[0,1]$, holding the baseline parameters $\mu = 0.4$, $s = 0.05$, and $s_0 = 0.1$ fixed (Figure 4). The three business models (the first-best two-part tariff, usage pricing, and subscription pricing) carry over unchanged. Figure 4 traces the equilibrium threshold and profit against the compute cost c as the learning rate η rises.

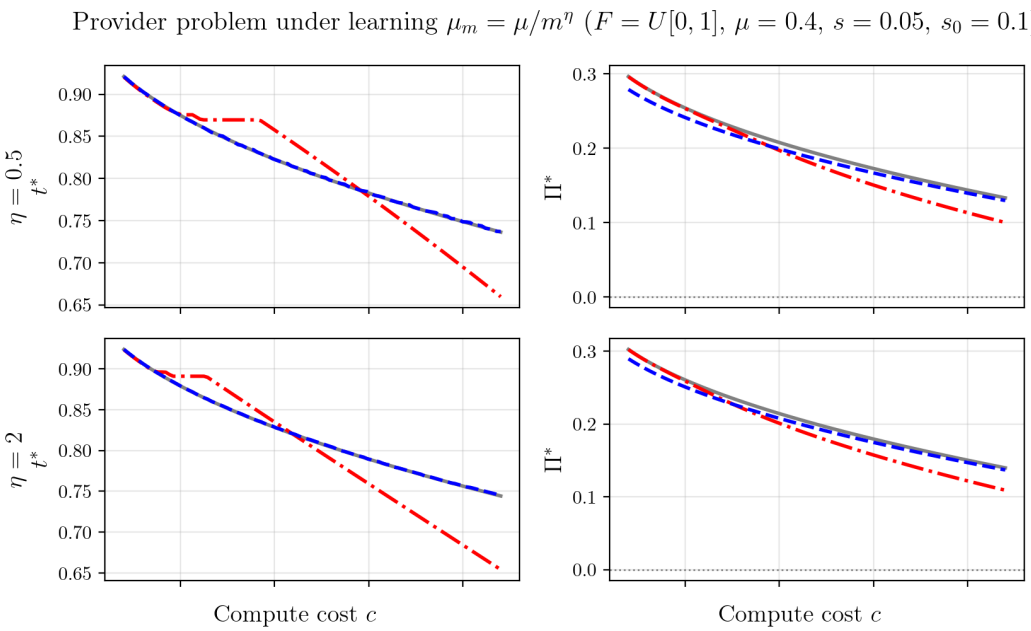

Figure 4: Equilibrium threshold t^* (left) and profit Π^* (right) versus compute cost c under the first-best, usage pricing, and subscription pricing, at learning rate $\eta \in \{0.5, 2\}$. The parameters are the same as Figure 1.

Two patterns stand out. First, learning raises profits Π^* across all three models. Intuitively, a lower effective noise along the search path raises the value of AI-delegated search (a higher r_1), so the profit panels shift upward as the learning rate η grows. Nonetheless, the profit comparison pattern between subscription and usage pricing remains the same as in Proposition 4.

Second, turning to the equilibrium threshold, usage pricing remains aligned with the efficient threshold t_{eff}. By contrast, learning reshapes the threshold distortions under subscription pricing. It shrinks the lower-tail range ($c < \underline{c}$, downward distortion) but expands the upper-tail range ($c > \bar{c}$, also a downward distortion). Intuitively, recall that the provider has two cost-control levers: lowering t to reduce the per-query compute cost, or adjusting t to curb the expected number of queries α. Under learning, acceptance probabilities rise along the path of their own accord, so the expected number of queries falls. Hence, controlling the number of queries becomes less pressing for the provider. In the upper-tail range, where low acceptance prevails, this implies more downward distortion (absent learning, the provider had to keep t not too low so as to prevent inducing excessive queries α); in the lower-tail range, where high acceptance prevails, Lemma 2 implies the reverse holds.

6.2 Blind acceptance of AI recommendations

The baseline model assumes that consumers inspect each AI recommendation before accepting it. In practice, however, some consumers act on an AI's recommendation without further inspection, e.g., accepting the answer returned by AI agents without clicking through to the underlying sources. This behavior attenuates the outer querying margin that creates the subscription distortion. We therefore extend the baseline monopoly model to allow for this possibility.

Consumer problem. Upon receiving a recommendation from the AI agent, the consumer now has three choices:

- *Blind-accept:* take the current recommendation without inspection.
- *Inspect:* pay inspection cost s to learn the realized match value of the current recommendation, then either accept it or query again, exactly as in the baseline.
- *Re-query:* query the AI agent for a new recommendation without inspecting the current recommendation.

In this environment, every step of a consumer's querying sequence involves two separate decisions. Under *usage pricing*, the consumer first decides whether to pay the per-query fee p to obtain a recommendation. Then, the consumer *separately* decides whether to blindly accept, additionally pay s to inspect the recommendation, or re-query for a new recommendation. Under *subscription pricing*, the same applies except $p = 0$. All other primitives are as in the baseline, and we maintain Assumption A throughout.

Define the *blind-accept value*

$$B(t) \equiv \int_{\underline{v}}^{\bar{v}} v \, dG(v|t), \tag{10}$$

i.e., the expected match value of an unseen recommendation. Because every recommendation is drawn from the same distribution $G(\cdot|t)$, a re-query offers nothing beyond the current blind-accept-or-inspect choice, so re-querying is dominated. The same stationarity makes the comparison between the blind-accept value and the continuation value of inspection identical at every instance of the querying sequence. Hence a consumer either blind-accepts the first recommendation, or queries and inspects every recommendation as in the baseline model. Therefore, the consumer's expected utility from signing up is

$$U(t,p) \;=\; \max\Big\{ \underbrace{B(t)-p}_{\text{blind-accept}}, \quad \underbrace{r(t,p)}_{\text{inspect (as in baseline)}} \Big\}, \tag{11}$$

generalizing the baseline expression $U = r(t,p)$. The participation constraint is $U(t,p) - T \geq v_0$, with $T = 0$ under usage pricing and $p = 0$ under subscription pricing as before.

Provider problems and outcomes. We use the first-best benchmark to illustrate how the possibility of blind acceptance affects the analysis. At the efficient marginal-cost pricing, total welfare under each consumer strategy (*blind-accept* or *inspect*) is, respectively,

$$\mathrm{TW}^{\mathrm{BA}}(t) \;=\; B(t) - \frac{c}{1-F(t)}, \qquad \mathrm{TW}^{\mathrm{I}}(t) \;=\; r(t, \tfrac{c}{1-F(t)}),$$

with the latter coinciding with the baseline welfare object. Under Assumption A, both functions are quasi-concave and *share the same maximizer* t_{eff} defined in (3), implying the efficient threshold is unaffected by allowing blind acceptance. Whether blind acceptance arises in equilibrium therefore depends on a direct comparison of the two welfare values at t_{eff}. To state the result, recall $r_{\text{eff}} = r(t_{\text{eff}}, \frac{c}{1-F(t_{\text{eff}})})$ using (3).

Lemma 4. *Under Assumption A, the first-best threshold remains* t_{eff} *defined in* (3)*. In the first-*

best outcome, consumers inspect every recommendation if

$$s \;<\; \mu \int_{\underline{v}}^{r_{\text{eff}}} F(v_i)\, dv_i;$$

and blindly accept the first recommendation otherwise.

Lemma 4 says that our baseline analysis carries over as long as consumers' cost of inspecting AI recommendations is not too high. The cutoff condition has a clean interpretation. Recall the noise event arises with probability μ, and conditional on noise the ex-ante distribution of v is F, so $\mu \int_{\underline{v}}^{r_{\text{eff}}} F(v)\, dv$ captures the expected gain from rejecting noise-driven recommendations whose realizations fall below r_{eff}. When this exceeds s, inspection pays for itself; otherwise blind acceptance is efficient. Intuitively, inspection is valuable when the consumer is selective. A high r_{eff} means the consumer finds it worthwhile to pay s for the option to inspect and reject low-value realizations.

The same dichotomy carries through to usage pricing and subscription pricing. In each case, the provider's optimization decomposes into two branches indexed by whether blind acceptance or inspection occurs in response to the provider's choices. The blind branch always achieves its unconstrained maximum at t_{eff}; the inspect branch coincides with the baseline analysis (Propositions 2 and 3). Whichever branch prevails in equilibrium is determined by which yields the provider higher profit. We relegate the formal analytical proofs and numerical illustrations to Online Appendix C.2. Two main qualitative insights emerge from our analysis with blind acceptance.

First, the efficient benchmark again entails either a high-acceptance or a blind-acceptance regime, both of which mitigate the double incidence of inspection costs. In the baseline, the efficient outcome recommends options that the consumer is expected to accept, so the consumer's inspection cost s is rarely incurred more than once before acceptance (Proposition 1). With blind acceptance, when s is sufficiently large relative to signal noisiness μ, the efficient outcome avoids the inspection cost s altogether. Hence, in this extension, the possibility of blind acceptance is *efficiency-improving*.

Second, the possibility of blind acceptance partially mitigates distortions of the baseline subscription model. In the baseline, the subscription threshold t_{subs} generally differs from t_{eff} and exhibits a double-crossing distortion structure (Proposition 3). With blind acceptance, the distortion vanishes whenever the equilibrium outcome falls on the blind-acceptance branch. This parallels the role of query caps in Section 4.2, in that both limit the number of consumer queries on the equilibrium path.

6.3 Sponsored options: transparency and opacity

We now consider a model of advertising where some options are sponsored by advertisers, so that the AI agent's recommendations can be either organic or sponsored. We ask how the presence of advertising revenues changes the provider's strategies and how *disclosure* of the recommendation type affects the provider. When the consumer can see whether a recommendation is sponsored, she can let her stopping decision depend on that type; when she cannot, she must treat every recommendation alike.

We build on the blind-acceptance model of Section 6.2 and, departing from the maintained

assumption $\mu > 0$, set $\mu = 0$ to isolate the disclosure channel.[24] The AI agent maintains an organic pool O and a sponsored pool S, screened with thresholds t_O and t_S. Both pools share the baseline value distribution. Conditional on type $j \in \{O, S\}$, the displayed recommendation is drawn from

$$G_j(v|t_j) = \frac{F(v) - F(t_j)}{1 - F(t_j)} \cdot \mathbf{1}_{\{v \geq t_j\}}.$$

At each query, the agent displays a sponsored recommendation with probability $\lambda \in [0, 1]$ and an organic recommendation with probability $1 - \lambda$, independently across queries. We treat λ as a primitive, capturing ad-load commitments. Only the displayed pool is screened by the AI agent, so the expected cost per displayed organic and per displayed sponsored recommendation is

$$\kappa_O \equiv \frac{c}{1 - F(t_O)}, \qquad \kappa_S \equiv \frac{c}{1 - F(t_S)}.$$

A sponsored recommendation earns a per-display impression ad revenue $A \geq 0$ and, additionally, a conversion ad revenue $a \geq 0$ if accepted by the consumer.

We compare two regimes. Under *opacity*, consumers do not observe the recommendation type and face a single mixture of the two pools; the equilibrium outcome parallels Section 6.2, in which re-querying never arises. Under *transparency* (equivalently, disclosure), consumers observe the type of recommendation before deciding, and can act differently in response to organic and sponsored recommendations. For instance, a consumer can re-query after one type while blindly accepting or inspecting the other. Online Appendix C.3 sets out the Bellman equation characterizing the consumer's reservation value, and the provider's profit. We then solve the model numerically and compare the equilibrium outcomes across opacity and transparency regimes.

Numerical finding: transparency vs. opacity. We present a set of numerical results based on $F \sim U[0, 1]$, $s = 0.01$, $c = 0.005$, and $v_0 = 0.65$. We vary $\lambda \in [0.10, 0.70]$ jointly with either impression revenue $A \in [0, 0.20]$ at $a = 0$, or conversion revenue $a \in [0, 0.20]$ at $A = 0$.

We plot the profit gain from sponsored-option disclosure, $\Pi^{\text{trp}} - \Pi^{\text{opq}}$, under subscription pricing (Figure 5) and usage pricing (Figure 6) respectively. Labels indicate, under the transparency regime, the consumer's actions following an organic and a sponsored recommendation, respectively: BA denotes blind acceptance, Q denotes re-querying, and I denotes inspection. For example, Q-I indicates that consumers re-query (inspect) upon seeing organic (sponsored) recommendations.

Our first finding in this extension is that transparency creates a "routing benefit". It allows the provider to assign distinct roles to the two recommendation pools within the querying process. This is clearest under subscription pricing (Figure 5). With impression revenue (first panel), transparency enables an attract-and-switch ad-exposure loop. One pool has a high enough threshold to ensure consumers join the querying sequence at all, while sponsored recommendations are mixed in to generate revenue. With conversion revenue (second panel), transparency lets the provider *tilt* consumers toward accepting sponsored options, by setting a relatively low inner-search threshold for organic recommendations and a high threshold for sponsored ones, which makes the latter more attractive. These routing benefits are absent under opacity because consumers apply a single

[24] In the baseline model without blind acceptance, options are pure search goods and the consumer decides only after observing the value realization. Consequently, disclosure has no effect. Moreover, the baseline analysis extends immediately to $\mu = 0$ with weak inequalities.

stopping rule to the pooled recommendation distribution.

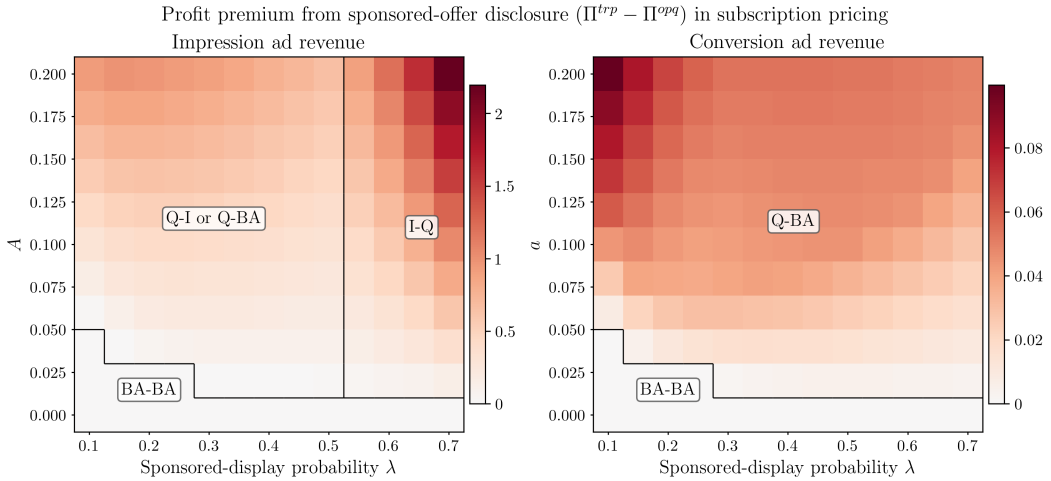

Figure 5: Profit premium from sponsored-option transparency under *subscription pricing*. Panels plot $\Pi^{\text{trp}} - \Pi^{\text{opq}}$ over (λ, A) at $a = 0$ (left) or over (λ, a) at $A = 0$ (right). Labels indicate, under transparency regime, consumer action after organic and sponsored recommendations, respectively.

The same routing benefit, in principle, exists under the usage pricing model (Figure 6), but numerical results suggest its magnitude is much smaller. With impression revenue (first panel) the transparency premium is essentially zero. With conversion revenue (second panel) transparency improves profit only when both a and λ are large, and even then the order of magnitude is smaller compared to the subscription pricing model. The reason is that the per-query fee under usage pricing already gives the provider an instrument for controlling the query path, and so the routing benefit of disclosure is relatively small.

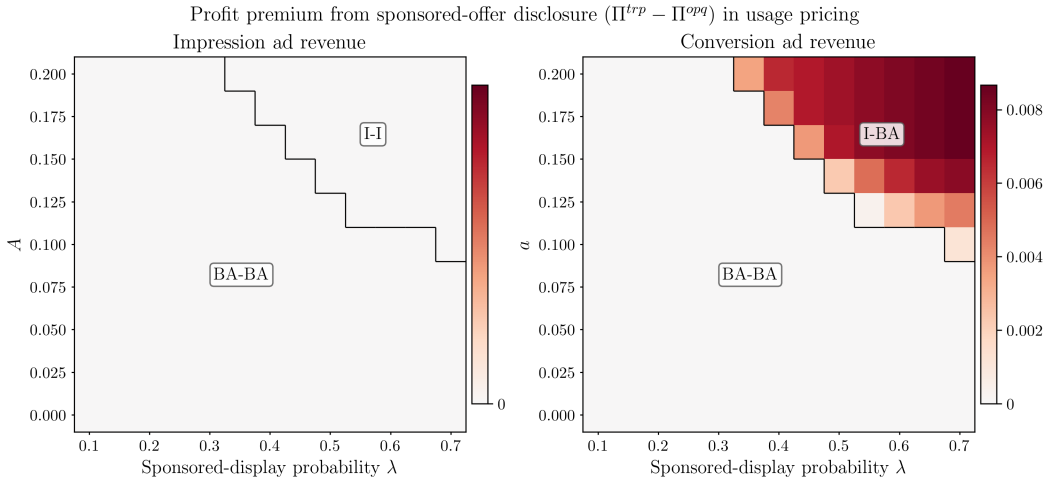

Figure 6: Profit premium from sponsored-option transparency under *usage pricing*. Panel layout and regime labels as in Figure 5.

The second finding is that a subscription provider stands to gain more from introducing impression-ad revenue than a usage-pricing provider (comparing the first panels of Figures 5 and 6). Intuitively, such revenue monetizes the excess query volume that subscription pricing tends to generate, turning its disadvantage into an advantage.[25]

[25] In Online Appendix C.4, we analyze a simpler, reduced-form model of advertising in which impression revenue

In summary, this extension shows that introducing a pool of sponsored recommendations does not merely add a revenue stream; it also changes how the provider designs recommendation thresholds and manages consumer querying. Sponsored-option disclosure induces type-contingent querying behavior that is unavailable under opacity, thereby raising provider profit. Nonetheless, a higher provider profit under transparency need not imply higher consumer surplus. Thus, the extension does not claim transparency is socially desirable in general; rather, transparency can be privately valuable as a way for providers to influence consumers' querying behavior.

7 Discussion and conclusions

7.1 Managerial implications

In this section, we discuss managerial implications and emphasize how they differ from standard search, recommendation, and quality-investment settings (e.g., search engines).

Managerial Insight 1. *AI providers should account for search-quality design when choosing a business model.*

A managerial implication is that pricing is not only a revenue instrument. In AI-delegated search, the business model shapes consumers' outer querying margin, which in turn changes the provider's optimal inner-search threshold and total compute cost. This point is most visible in the contrast between usage and subscription pricing, as shown in Propositions 2-3. Therefore, a provider that first chooses a search-quality design and only later chooses between subscription and usage pricing may misdiagnose the true cost of serving users. The design problem and the monetization problem are jointly determined.

Managerial Insight 2. *Efficient AI-delegated search should minimize unnecessary follow-up queries, not simply maximize apparent recommendation quality.*

Unlike traditional search engines, AI-delegated search involves a double incidence of search costs. The provider pays compute cost to screen options, while the user may still pay inspection cost to verify the recommendation. In the efficient benchmark, the AI agent recommends options that, to the best of its information, the user is willing to accept. With precise signals, this can mean a single high-quality recommendation; with noisier signals, it means more frequent user check-ins, but still with the goal of avoiding unnecessary follow-up queries.[26]

Managerial Insight 3. *The provider's cost-control strategy under subscription can involve either excessively low or excessively high inner-search quality.*

This insight highlights another departure from standard "quality investment" settings, where total cost is typically reduced by lowering quality. In AI-delegated search, however, total cost is driven by user querying behavior, which itself responds to the search quality. The provider can thus

lowers the provider's effective per-query marginal cost from $\frac{c}{1-F(t)}$ to $\frac{c}{1-F(t)} - A$, where $A \geq 0$ denotes the per-query ad revenue. We analytically prove that a subscription-pricing provider has a stronger incentive than a usage-pricing provider to initiate impression-ad monetization.

[26] Anecdotally, compared with users sent to a website by Google, those sent by ChatGPT tend to spend more time on a site, view more pages, and are more likely to complete a transaction (Wall Street Journal, 2026), suggesting AI-delegated search involves fewer follow-up queries than traditional search engines.

control total cost in two opposing ways: raise the inner-search threshold so that recommendations are accepted early; or lower it so that each query is cheap but increases follow-up queries. This explains the paper's double-crossing distortion result under subscription pricing.

The first route, paying a higher cost per query in exchange for far fewer queries, is the counterintuitive one, because better service is delivered at lower total cost. The same logic has been observed in agentic-AI deployment. For example, coding practitioners report that a top-tier model (e.g., Anthropic's Opus class) reaches a working solution in one or two passes, while a cheaper model needs several rounds of revision, so the premium model's cumulative bill can end up lower despite a higher per-token price.[27] In our model, the AI provider plays the analogous role. A higher search threshold can reduce the number of queries and thus lower the total compute or API cost incurred by the provider.[28]

Managerial Insight 4. *Usage pricing is more profitable when competition is intense or when the cost of AI evaluation is high; subscription pricing is more profitable in the opposite cases.*

The business-model profit comparison follows from the same alignment mechanism. Usage pricing is more attractive when the benefit of aligning marginal queries is large, i.e., compute costs are high, tasks are compute-intensive, or users have strong outside options. This is consistent with the prevalence of usage-based pricing in developer-facing API markets and other settings where providers compete on marginal value and serve computationally intensive tasks (Bergemann et al., 2026). Conversely, subscription pricing is more attractive when the compute costs are low or when the competing providers are highly horizontally differentiated, e.g., in the consumer-facing market.

Managerial Insight 5. *Usage caps and advertising mitigate, but do not fully resolve, the over-querying inefficiencies associated with subscription pricing.*

Query caps limit exposure to excessive use while preserving the consumer-facing simplicity of a flat plan. However, they are not a perfect substitute for usage pricing. Because realized query demand is stochastic, any fixed cap either binds when additional search would be valuable or fails to bind when querying is excessive. In contrast, usage pricing aligns marginal incentives in every realized query state.

Advertising changes the same trade-off in the opposite direction. Impression revenue monetizes the excess query volume generated by subscription pricing, giving subscription providers stronger incentives to introduce advertising than usage-pricing providers. Meanwhile, in a richer setting where the AI agent draws options from organic and sponsored pools, sponsored-option transparency enables a routing instrument whereby the provider can assign different roles to organic and sponsored recommendations to influence consumers' query path.

7.2 Policy implications

Next, we discuss how the results of our study inform policymakers regarding platform and AI governance.

[27] Anthropic (2025) reports that Claude Opus 4.5 matches the prior Sonnet's best SWE-bench score using roughly 76% fewer output tokens; industry commentary on long-running agents emphasizes "cost-per-outcome" rather than cost-per-token as the relevant metric (Caylent, 2026).

[28] Nonetheless, the second route of lowering the threshold to economize on per-query costs also appears in practice. For example, industry reporting suggests that leading providers respond to compute and cost constraints by reducing reasoning depth or effort levels (VentureBeat, 2026; InfoWorld, 2026).

Policy Insight 1. *Flat-rate AI access can create hidden quality distortions, even when it appears consumer-friendly.*

From the regulatory perspective, subscription pricing protects consumers from unexpected marginal charges, but it also affords users a zero marginal price per query. This practice can induce users to over-query because they do not internalize the provider's compute or API cost. The provider then responds by changing the inner-search threshold, which can make search quality inefficiently high or low depending on compute-cost conditions. Therefore, policies encouraging universal, flat-rate AI access, such as the concept of "Universal Basic AI Access" (Špecián and Špeciánová, 2025), should not evaluate consumer protection only through price predictability. Rather, regulators should ask whether flat pricing changes provider incentives in ways that degrade recommendation quality, encourage shallow search, or produce unnecessary repeat queries. A flat-rate plan may be affordable at the point of purchase but costly in terms of distorted AI outputs.

Policy Insight 2. *Usage-based pricing should not be treated as inherently harmful. It can be an efficiency-enhancing pricing instrument.*

Conventional wisdom usually views usage-based (i.e., per-query or per-token) pricing as less consumer-friendly than subscriptions (Yun and Suk, 2022). In this study, we show that, under usage pricing, consumers face a marginal price for each query, and the provider chooses the efficient search-quality threshold. In contrast, subscription pricing can separate the user's querying margin from the provider's compute-cost margin. Hence, from the policy perspective, usage-based pricing, compute credits, or two-part tariffs may be socially useful when they make users account for costly AI search effort. Consumer-protection regulations should not prohibit marginal pricing altogether. Instead, they could focus on aspects such as transparency and fairness, which include clear units, clear overage rules, spend caps, and understandable cost estimates.

Policy Insight 3. *AI search-quality regulation should be outcome-based rather than input-based.*

Our study demonstrates that efficient AI-delegated search minimizes unnecessary follow-up queries while balancing informational gain against compute cost. A higher threshold can sometimes reduce total cost by producing recommendations users accept earlier. In this regard, when formulating regulations, policymakers should consider outcome-based metrics such as repeat-query rates, correction rates, user satisfaction after acceptance, compute per solved task, and the share of sponsored versus organic accepted recommendations. Meanwhile, conventional measures, such as latency, tokens used, or the number of model calls, may not be effective because they do not capture whether the AI search process truly helped users efficiently reach the desired outcome.

Policy Insight 4. *Sponsored AI recommendations require disclosure, but disclosure alone is not enough.*

Our analysis of sponsored options shows that transparency lets consumers condition their stopping behavior on whether a recommendation is organic or sponsored. Such a practice can raise provider profit, especially under subscription pricing, but it does not necessarily benefit consumers. Advertising also monetizes the excess query volume that subscriptions tend to generate. Accordingly, policy should require clear labeling of sponsored recommendations, but it should also examine

the routing logic behind AI suggestions. For example, regulators could verify whether sponsored results are shown earlier, whether organic and sponsored thresholds differ, whether users are nudged into additional queries, and whether disclosure changes stopping behavior in ways that benefit the provider more than the user.

7.3 Concluding remarks

In AI-delegated search, the user controls the outer querying margin, while the provider controls the inner-search threshold and bears the compute cost of internal searches by its AI agents. Business models determine whether these margins are aligned and, in turn, shape providers' search-quality design choices.

Our framework of AI-delegated search offers a tractable setting for analyzing markets in which agents act as interactive search intermediaries rather than passive information channels, and highlights how such settings differ from traditional search. It can be a building block for other research themes on markets with AI agents. First, the framework can be used to study how AI-delegated search affects the prevailing ad-funded publisher ecosystem, which has long been supported by traffic referrals from traditional search engines such as Google. Second, it can be extended dynamically to allow for a data-feedback loop between AI and consumers, generating richer intertemporal trade-offs. Third, our search environment can be extended to allow for heterogeneous usage intensity and the analysis of optimal menu pricing using mechanism-design techniques.

Empirically, our comparative statics in the cost parameter c imply a dynamic prediction about business-model transitions. The rapid decline in per-token inference costs documented by Demirer et al. (2025) would, on its own, push providers toward subscription pricing as the cost of misalignment shrinks. However, our model emphasizes that the relevant object is instead the effective compute cost of completing a user query or task. When agentic and long-horizon workloads raise task-level compute costs, our result predicts a shift toward usage pricing or tighter caps. A vivid recent illustration is GitHub's June 2026 transition of Copilot from a flat-rate plan to usage-based billing (GitHub, 2026): "*Today, a quick chat question and a multi-hour autonomous coding session can cost the user the same amount. GitHub has absorbed much of the escalating inference cost behind that usage, but the current premium request model is no longer sustainable.*"[29] Thus, as AI agents mature and become integrated into everyday use, our prediction is a broader adoption of usage-based pricing: even consumer-facing AI services that began as flat-rate subscriptions may increasingly incorporate usage-based charges, compute credits, or binding usage limits.[30]

References

Anthropic (2025) "Introducing Claude Opus 4.5," Available at anthropic.com/news/claude-opus-4-5.

[29] Google (2026) has made a similar announcement in May 2026 for its Gemini app: it has shifted to a "credit-based quota system" in which usage limits scale with the complexity of the prompt and the length of the chat; Google AI Pro and Ultra subscribers can purchase pay-as-you-go top-up credits once these caps are reached.

[30] Recent product developments suggest that agentic AI is moving beyond developer-facing tools and into general-purpose workflows for non-technical users. Examples include Anthropic's Claude Cowork, OpenAI's Codex desktop app, and open-source personal-agent projects such as OpenClaw.

Armstrong, Mark and John Vickers (2010) "A Model of Delegated Project Choice," *Econometrica*, 78 (1), 213–244.

Armstrong, Mark, John Vickers, and Jidong Zhou (2009) "Prominence and Consumer Search," *RAND Journal of Economics*, 40 (2), 209–233.

Armstrong, Mark and Jidong Zhou (2011) "Paying for Prominence," *Economic Journal*, 121 (556), F368–F395.

Athey, Susan and Glenn Ellison (2011) "Position Auctions with Consumer Search," *Quarterly Journal of Economics*, 126 (3), 1213–1270.

Bagnoli, Mark and Theodore C. Bergstrom (2005) "Log-concave probability and its applications," *Economic Theory*, 26 (2), 445–469.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin (2026) "Menu Pricing of Large Language Models," arXiv preprint arXiv:2502.07736.

Bick, Alexander, Adam Blandin, and David J Deming (2026) "The rapid adoption of generative AI," *Management Science* (Forthcoming).

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello (2020) "Artificial Intelligence, Algorithmic Pricing, and Collusion," *American Economic Review*, 110 (10), 3267–3297.

Caylent (2026) "Claude Opus 4.7 Deep Dive: Capabilities, Migration, and the New Economics of Long-Running Agents," Available at caylent.com/blog.

Chen, Yongmin and Tianle Zhang (2018) "Intermediaries and Consumer Search," *International Journal of Industrial Organization*, 57, 255–277.

De Cornière, Alexandre and Greg Taylor (2014) "Integration and Search Engine Bias," *RAND Journal of Economics*, 45 (3), 576–597.

——— (2019) "A Model of Biased Intermediation," *RAND Journal of Economics*, 50 (4), 854–882.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis, and Sida Peng (2025) "The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs," Working Paper34608, National Bureau of Economic Research.

Eliaz, Kfir and Ran Spiegler (2011) "A Simple Model of Search Engine Pricing," *Economic Journal*, 121 (556), F329–F339.

Erat, Sanjiv and Vish Krishnan (2012) "Managing Delegated Search Over Design Spaces," *Management Science*, 58 (3), 606–623.

Fish, Sara, Yannai A. Gonczarowski, and Ran I. Shorrer (2024) "Algorithmic Collusion by Large Language Models," arXiv preprint arXiv:2404.00806.

Guo, Jiafeng, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng (2022) "Semantic Models for the First-stage Retrieval: A Comprehensive Review," *ACM Transactions on Information Systems*, 40 (4), 1–42.

Guo, Liang (2023) "Gathering Information before Negotiation," *Management Science*, 69 (1), 200–219.

Hadfield, Gillian K. and Andrew Koh (2025) "An Economy of AI Agents," arXiv preprint arXiv:2509.01063.

Hagiu, Andrei and Julian Wright (2020) "Platforms and the Exploration of New Products," *Management Science*, 66 (4), 1527–1543.

Immorlica, Nicole, Brendan Lucier, and Aleksandrs Slivkins (2024) "Generative ai as economic agents," *ACM SIGecom Exchanges*, 22 (1), 93–109.

Inderst, Roman and Marco Ottaviani (2012) "Competition through Commissions and Kickbacks," *American Economic Review*, 102 (2), 780–809.

InfoWorld (2026) "Enterprise Developers Question Claude Code's Reliability for Complex Engineering," https://www.infoworld.com/article/4154973/enterprise-developers-question-claude-codes-reliability-for-complex-engineering.html.

Janssen, Maarten C. W. and Cole Williams (2024) "Influencing Search," *The RAND Journal of Economics*, 55 (3), 442–462.

Johnson, Justin P. and David P. Myatt (2006) "On the Simple Economics of Advertising, Marketing, and Product Design," *American Economic Review*, 96 (3), 756–784.

Kaiser, Carolin, Jakob Kaiser, René Schallner, and Sabrina Schneider (2025) "A New Era of Online Search? A Large-Scale Study of User Behavior and Personal Preferences during Practical Search Tasks with Generative AI versus Traditional Search Engines," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*.

Karle, Heiko, Marcel Preuss, and Markus Reisinger (2026) "Selling on Recommender Platforms: Demand Boost versus Customer Migration," SSRN Working Paper 6153966.

Ke, Tony, Song Lin, and Michelle Y. Lu (forthcoming) "Information Design of Online Platforms," *Management Science*.

Lewis, Tracy R. (2012) "A Theory of Delegated Search for the Best Alternative," *RAND Journal of Economics*, 43 (3), 391–416.

Lewis, Tracy R. and David E. M. Sappington (1994) "Supplying Information to Facilitate Price Discrimination," *International Economic Review*, 35 (2), 309–327.

Liang, Yidong, Zhijing Wu, Fan Zhang, Dandan Song, and Heyan Huang (2025) "How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, 634–644.

Mahmood, Rafid (2024) "Pricing and Competition for Generative AI," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Ng, Robin and Michael Wessel (2026) "AI Overview or Overreach? Google's Strategic Deployment of Generative AI in Search," Discussion Paper742, Collaborative Research Center Transregio 224 (CRC TR 224).

Prelec, Drazen and George Loewenstein (1998) "The Red and the Black: Mental Accounting of Savings and Debt," *Marketing Science*, 17 (1), 4–28.

Quispe-Torreblanca, Edika G., Neil Stewart, John Gathergood, and George Loewenstein (2019) "The Red, the Black, and the Plastic: Paying Down Credit Card Debt for Hotels, Not Sofas," *Management Science*, 65 (11), 5392–5410.

Špecián, Petr and Jitka Špeciánová (2025) "Universal Basic AI Access: Countering the Digital Divide," *Acta Informatica Pragensia*, 14 (2), 272–281.

Teh, Tat-How and Julian Wright (2022) "Intermediation and Steering: Competition in Prices and Commissions," *American Economic Journal: Microeconomics*, 14 (2), 281–321.

Ulbricht, Robert (2016) "Optimal Delegated Search with Adverse Selection and Moral Hazard," *Theoretical Economics*, 11 (1), 253–278.

VentureBeat (2026) "Is Anthropic "Nerfing" Claude? Users Increasingly Report Performance Degradation as Leaders Push Back," https://venturebeat.com/technology/is-anthropic-nerfing-claude-users-increasingly-report-performance.

Wall Street Journal (2026) "What Is GEO? The New SEO Skill Every Business Needs," https://www.wsj.com/tech/ai/ai-what-is-geo-aeo-5c452500.

Weitzman, Martin L. (1979) "Optimal Search for the Best Alternative," *Econometrica*, 47 (3), 641–654.

Yun, Sejeong and Kwanho Suk (2022) "Consumer preference for pay-per-use service tariffs: the roles of mental accounting," *Journal of the Academy of Marketing Science*, 50 (5), 1111–1124.

Zhong, Zemin (2023) "Platform Search Design: The Roles of Precision and Price," *Marketing Science*, 42 (2), 293–313.

Zorc, Saša, Ilia Tsetlin, Sameer Hasija, and Stephen E. Chick (2025) "The When and How of Delegated Search," *Operations Research*, 73 (1), 461–482.

Zou, Tianxin and Bo Zhou (2025) "Self-Preferencing and Search Neutrality in Online Retail Platforms," *Management Science*, 71 (5), 4087–4107.

8 Appendix: Proofs

Proof. **(Lemma 1).** We note that a higher t implies distribution $G(\cdot|t)$ becomes higher in the sense of first-order stochastic dominance. Therefore $\Phi(v|t)$ is increasing in t. The result follows immediately from the definition (2) and the fact that $\Phi(v|t)$ is decreasing in v. □

Proof. **(Lemma 2).** We rewrite $\Phi(v|t)$ as

$$\Phi(v|t) = \begin{cases} \dfrac{1-\mu F(t)}{1-F(t)} \displaystyle\int_v^{\bar{v}} (1-F(v_i))\, dv_i & \text{if } v \geq t \quad \text{(branch LA)} \\ \displaystyle\int_v^t [1-\mu F(v_i)]\, dv_i + \dfrac{1-\mu F(t)}{1-F(t)} \displaystyle\int_t^{\bar{v}} [1-F(v_i)]\, dv_i & \text{if } v \leq t \quad \text{(branch HA)} \end{cases} \tag{12}$$

We first construct the boundary threshold $\hat{t}$. Using the first row of (12), we define function

$$\Lambda(t) \equiv (1-\mu F(t)) \cdot \text{MRL}(t) \tag{13}$$

which is decreasing by (1). Observe that $\Lambda(\bar{v}) = 0$. By construction, $\Lambda(t) < s+p \Leftrightarrow \Phi(t|t) < s+p \Leftrightarrow t > r(t,p)$. If $\Lambda(\underline{v}) < s+p$, then $t > r(t,p)$ for all $t \in [\underline{v}, \bar{v}]$ and so we define $\hat{t} = \underline{v}$ without loss of generality. If $\Lambda(\underline{v}) \geq s+p$, then the intermediate value theorem implies the existence of a unique cutoff $\hat{t} \in [\underline{v}, \bar{v}]$ that solves $\Lambda(\hat{t}) = s+p$, such that $t < \hat{t} \Leftrightarrow t < r(t,p)$.

If $t < \hat{t}$, the low-acceptance (LA) regime implies $\alpha(t,p) = \frac{1}{s+p}\text{MRL}(r)$, which is decreasing in r,

$$\begin{aligned} \frac{\partial}{\partial p} r(t,p) &= \frac{1}{\partial \Phi(v|t)/\partial v}|_{v=r} = -\left(\frac{1-F(t)}{1-F(r)}\right) \frac{1}{1-\mu F(t)} < 0 \\ \frac{\partial}{\partial t} r(t,p) &= -\frac{\partial \Phi(v|t)/\partial t}{\partial \Phi(v|t)/\partial v}|_{v=r} = \frac{s+p}{[1-\mu F(t)]^2} \cdot \frac{(1-\mu) f(t)}{1-F(r)} > 0. \end{aligned}$$

where the simplification used the definition of $\Phi(r(t,p)|t) = s+p$.

If $t \geq \hat{t}$ the high-acceptance (HA) regime implies $\alpha(t,p) = \frac{1}{1-\mu F(r)}$ is increasing in r, and the derivatives are (with a slight abuse of notation, the derivative notation is understood as the right derivative when evaluated at the kink point $t = \hat{t}$):

$$\begin{aligned} \frac{\partial}{\partial p} r(t,p) &= \frac{1}{\partial \Phi(v|t)/\partial v}|_{v=r} = -\frac{1}{1-\mu F(r)} < 0 \\ \frac{\partial}{\partial t} r(t,p) &= -\frac{\partial \Phi(v|t)/\partial t}{\partial \Phi(v|t)/\partial v}|_{v=r} = \frac{1}{1-\mu F(r)} \cdot \text{MRL}(t) \cdot \frac{(1-\mu) f(t)}{1-F(t)} \in (0,1]. \end{aligned} \tag{14}$$

where we used DMRL, because $\text{MRL}'(t) = \text{MRL}(t)\frac{f(t)}{1-F(t)} - 1 \leq 0$. □

Proof. **(Proposition 1).** Using (13) in the proof of Lemma 2, we define $\hat{t}_{\text{eff}}$ as the unique solution to:

$$\Lambda(\hat{t}_{\text{eff}}) - s - \frac{c}{1-F(\hat{t}_{\text{eff}})} = 0 \tag{15}$$

if a solution exists, and define $\hat{t}_{\text{eff}} = \underline{v}$ otherwise. Then, $t < \hat{t}_{\text{eff}}$ if and only if $t < r(t, \frac{c}{1-F(t)})$.

Branch LA ($t < \hat{t}_{\text{eff}}$). Using the first case of (12), and derivatives in the proof of Lemma 2:

$$\begin{aligned}\frac{d}{dt}r(t, \frac{c}{1-F(t)}) &= \left[\frac{\partial r(t,p)}{\partial t} + \frac{\partial r(t,p)}{\partial p}\frac{f(t)c}{(1-F(t))^2}\right]_{p=\frac{c}{1-F(t)}} \\ &= \frac{f(t)}{[1-F(r)][1-\mu F(t)]^2}\cdot[(1-\mu)s - \mu c]\end{aligned}$$

Thus, the supremum for this branch occurs at $t = \underline{v}$ if $\mu c > (1-\mu)s$, and at $t = \hat{t}_{\text{eff}}$ if $\mu c < (1-\mu)s$; in the boundary case $\mu c = (1-\mu)s$ the derivative vanishes, so $r(t, \frac{c}{1-F(t)})$ is constant on $[\underline{v}, \hat{t}_{\text{eff}}]$.

Branch HA ($t \geq \hat{t}_{\text{eff}}$). Using the second case of (12) and derivatives in the proof of Lemma 2:

$$\frac{d}{dt}r(t, \frac{c}{1-F(t)}) = \frac{f(t)}{[1-\mu F(r)][1-F(t)]^2}\cdot\left[(1-\mu)\int_t^{\bar{v}}[1-F(v_i)]\,dv_i - c\right]$$

which is single-crossing in t, implying the objective function $r(t, \frac{c}{1-F(t)})$ is strictly quasi-concave. The first-order condition is satisfied at t° that satisfies $\int_{t^\circ}^{\bar{v}}[1-F(v_i)]\,dv_i = \frac{c}{1-\mu}$. Then, plugging t° into the definition of $\hat{t}_{\text{eff}}$ in (15) and simplifying, we have $t^\circ \geq \hat{t}_{\text{eff}}$ if and only if

$$\frac{1-\mu F(t^\circ)}{1-F(t^\circ)}\cdot\frac{c}{1-\mu} - s - \frac{c}{1-F(t^\circ)} \leq 0$$

which holds if and only if $\mu c \leq (1-\mu)s$.

Combining both branches using continuity of $r(t, \frac{c}{1-F(t)})$ at $t = \hat{t}_{\text{eff}}$ proves the result: $t_{\text{eff}} = t^\circ$ if $\mu c < (1-\mu)s$, and $t_{\text{eff}} = \underline{v}$ if $\mu c > (1-\mu)s$. At the non-generic boundary case $\mu c = (1-\mu)s$, $t^\circ = \hat{t}_{\text{eff}}$ is the unique optimizer of the HA branch, whereas the LA-branch flatness noted above means $[\underline{v}, \hat{t}_{\text{eff}})$ are also optimal; without loss, we select $t_{\text{eff}} = \underline{v}$. □

Proof. **(Proposition 2).** Continuing from (4), we first fix target reservation value $u \in [v_0, r_{\text{eff}}]$, which is non-empty by Assumption A, and optimize with respect to t to show that the optimal threshold is independent of u. We then optimize with respect to u and verify that the optimum is $u_{\text{usage}} \in [v_0, r_{\text{eff}}]$.

Branch LA ($t < u$). Using the first case of (12) and $1 - G(v|t) = \frac{(1-\mu F(t))}{1-F(t)}(1-F(v))$ the profit becomes

$$\begin{aligned}\Pi_{\text{usage}}(t,u) &= \text{MRL}(u) - \frac{s[1-F(t)]+c}{(1-\mu F(t))(1-\mu F(u))}. \\ \frac{\partial}{\partial t}\Pi_{\text{usage}}(t,u) &= \frac{f(t)}{(1-\mu F(u))(1-\mu F(t))^2}\cdot[(1-\mu)s - \mu c] > 0\end{aligned}$$

where the inequality is implied by Assumption A. Therefore, profit is strictly increasing in t within this branch, so the Branch LA supremum is approached as $t \to u$.

Branch HA ($t \geq u$). Using the second case of (12), the profit simplifies to

$$\Pi_{\text{usage}}(t,u) = \frac{1}{1-\mu F(u)}\left(\Phi(u|t) - s - \frac{c}{1-F(t)}\right).$$

Differentiating,

$$\frac{\partial}{\partial t}\Pi_{\text{usage}}(t,u) = \frac{f(t)}{[1-\mu F(u)][1-F(t)]^2}\cdot\left((1-\mu)\int_t^{\bar{v}}[1-F(v_i)]\,dv_i - c\right),$$

which is strictly single-crossing in t, implying strict quasiconcavity of profit function in t. It has the exact same first-order condition as t_{eff}, which lies in Branch HA since $t_{\text{eff}} > r_{\text{eff}} \geq u$ (the initial supposition).

Combining both branches and invoking continuity of $\Pi_{\text{usage}}(t)$ at $t = u$, we conclude $t = t_{\text{eff}}$ (branch HA) is optimal and independent of u. For the comparative statics, (3) shows that $t_{\text{usage}} = t_{\text{eff}}$ is decreasing in c and μ. Next, substituting the optimal threshold $t = t_{\text{eff}}$ (with HA branch) and differentiating (4), the first-order condition for the optimal target utility is

$$\begin{aligned}\frac{d}{du}\Pi_{\text{usage}}(t_{\text{eff}},u) &= -1 + \left(\Phi(u|t_{\text{eff}}) - s - \frac{c}{1-F(t_{\text{eff}})}\right)\cdot\frac{\mu f(u)}{[1-\mu F(u)]^2}\\ &= -1 + \left(\int_u^{r_{\text{eff}}}[1-\mu F(v_i)]\,dv_i\right)\cdot\frac{\mu f(u)}{[1-\mu F(u)]^2} \quad = \quad 0. \end{aligned} \tag{16}$$

where we used the definition $\Phi(r_{\text{eff}}|t_{\text{eff}}) = s + \frac{c}{1-F(t_{\text{eff}})}$. The participation constraint $u \geq v_0$ binds if and only if $\frac{d\Pi_{\text{usage}}}{du}|_{u=v_0} \leq 0$, whereas $\frac{d\Pi_{\text{usage}}}{du}|_{u=r_{\text{eff}}} = -1 < 0$. So, $u_{\text{usage}} \in [v_0, r_{\text{eff}}]$ as claimed. □

Proof. **(Proposition 3).** Using (13) in the proof of Lemma 2, define $\hat{t}_{\text{subs}}$ as the unique solution to:

$$\Lambda(\hat{t}_{\text{subs}}) - s = 0. \tag{17}$$

The existence of $\hat{t}_{\text{subs}} \in (\underline{v}, \bar{v})$ follows from the intermediate value theorem because the boundary condition is satisfied: $(1-\mu F(\underline{v}))\cdot \text{MRL}(\underline{v}) = \mathbb{E}[v_i] - \underline{v} > s_0 \geq s$. Then, $t < \hat{t}_{\text{subs}} \Leftrightarrow t < r(t,0)$. In what follows, we sometimes write $r = r(t,0)$ and $\hat{t} = \hat{t}_{\text{subs}}$ to simplify notation.

Branch HA ($t \geq \hat{t}_{\text{subs}}$). Here $r(t,0) < t$, so the subscription profit is

$$\Pi_{\text{HA}}(t) = r(t,0) - v_0 - \frac{c}{(1-F(t))\big(1-\mu F(r)\big)}, \tag{18}$$

$$\Pi'_{\text{HA}}(t) = \frac{(1-\mu)f(t)\cdot\text{MRL}(t)}{(1-F(t))(1-\mu F(r))}\,\Gamma_{\text{HA}}(t), \tag{19}$$

where we used the derivative (14) from the proof of Lemma 2, and

$$\Gamma_{\text{HA}}(t) \equiv 1 - \underbrace{\frac{c\,\mu\,f(r)}{(1-F(t))(1-\mu F(r))^2}}_{\equiv\zeta(t)} - \frac{c}{(1-\mu)(1-F(t))\cdot\text{MRL}(t)}. \tag{20}$$

We claim $\zeta(t)$ is weakly increasing in t because its log-derivative is

$$\frac{d}{dt}\ln\zeta(t) = \frac{f(t)}{1-F(t)} + \frac{f'(r)}{f(r)}\frac{\partial r}{\partial t} + \frac{2\mu f(r)}{1-\mu F(r)}\frac{\partial r}{\partial t} \geq \frac{f(t)}{1-F(t)} - \frac{f(r)}{1-F(r)} \geq 0$$

where the first inequality is due to (1) and (14) and the second inequality is due to $r < t$ in this branch. Therefore (20) implies $\Gamma_{\text{HA}}(t)$ is strictly decreasing in t, implying Π_{HA} is strictly quasi-concave and has at most one interior solution t_{HA}, characterized by $\Gamma_{\text{HA}}(t_{\text{HA}}) = 0$.

Define c_{HA} as the value at which $\Gamma_{\text{HA}}(\hat{t}_{\text{subs}}) = 0$, i.e., $t_{\text{HA}} = \hat{t}_{\text{subs}}$. Solving the linear equation yields

$$c_{\text{HA}} = \frac{(1-\mu)(1-F(\hat{t}))\,\text{MRL}(\hat{t})}{1 + \dfrac{(1-\mu)\,\mu\,f(\hat{t})\,\text{MRL}(\hat{t})}{(1-\mu F(\hat{t}))^2}},$$

such that $t_{\text{HA}} > \hat{t}_{\text{subs}} \Leftrightarrow c < c_{\text{HA}}$.

Branch LA ($t < \hat{t}_{\text{subs}}$). Here $r(t,0) > t$, so the subscription profit is

$$\Pi_{\text{LA}}(t) = r(t,0) - v_0 - \frac{c}{(1-F(r))(1-\mu F(t))}.$$

$$\Pi'_{\text{LA}}(t) = \frac{s(1-\mu)\,f(t)}{(1-F(r))(1-\mu F(t))^2}\,\Gamma_{\text{LA}}(t),$$

where we used the derivative $\partial r/\partial t$ from the proof of Lemma 2, and

$$\Gamma_{\text{LA}}(t) \equiv 1 - \frac{c\mu}{s(1-\mu)} - \frac{c\,f(r)}{(1-F(r))^2(1-\mu F(t))}. \tag{21}$$

Observe that the RHS of (21) is strictly decreasing in t and r by (1) and that $r(t,0)$ is increasing in t by Lemma 1. Therefore $\Gamma_{\text{LA}}(t)$ is strictly decreasing in t, implying Π_{LA} is strictly quasi-concave and has at most one interior solution t_{LA}, characterized by $\Gamma_{\text{LA}}(t_{\text{LA}}) = 0$. Define c_{LA} as the value of c at which $\Gamma_{\text{LA}}(\hat{t}_{\text{subs}}) = 0$, i.e., $t_{\text{LA}} = \hat{t}_{\text{subs}}$. Solving the linear equation yields

$$c_{\text{LA}} = \frac{s(1-\mu)}{\mu + \dfrac{s(1-\mu)\,f(\hat{t})}{(1-F(\hat{t}))^2(1-\mu F(\hat{t}))}} < \frac{s(1-\mu)}{\mu},$$

such that $t_{\text{LA}} < \hat{t}_{\text{subs}} \Leftrightarrow c > c_{\text{LA}}$.

Combining branches. We first show $c_{\text{HA}} < c_{\text{LA}}$. We evaluate $\Gamma_{\text{LA}}(\hat{t})$ at $c = c_{\text{HA}}$ (equivalent to $\Gamma_{\text{HA}}(\hat{t}) = 0$). We denote the hazard rate as $\varphi(t) = \frac{f(t)}{1-F(t)}$ and write $b \equiv 1 - \mu F(\hat{t})$, and note $s = b\,\text{MRL}(\hat{t})$. Then,

$$\Gamma_{\text{LA}}(\hat{t})\big|_{c=c_{\text{HA}}} = \frac{(1-\mu)^2\,b\,\text{MRL}(\hat{t})}{b^2 + (1-\mu)\mu\,f(\hat{t})\text{MRL}(\hat{t})}\,\big[b - (1-\mu)\,\varphi(\hat{t})\text{MRL}(\hat{t})\big].$$

where $[b - (1-\mu)\,\varphi(\hat{t})\text{MRL}(\hat{t})] \geq \mu(1-F(\hat{t}))$ by decreasing MRL (i.e., $\varphi(\hat{t})\,\text{MRL}(\hat{t}) \leq 1$). So $\Gamma_{\text{LA}}(\hat{t})\big|_{c=c_{\text{HA}}} > 0$, which gives $c_{\text{LA}} > c_{\text{HA}}$. Then, we conclude

$$t_{\text{subs}} = \begin{cases} t_{\text{HA}} & \text{if } c < c_{\text{HA}}, \\ \hat{t}_{\text{subs}} & \text{if } c_{\text{HA}} \leq c \leq c_{\text{LA}}, \\ t_{\text{LA}} & \text{if } c > c_{\text{LA}}. \end{cases} \tag{22}$$

To see this, note for $c \in [c_{\text{HA}}, c_{\text{LA}}]$, neither branch has a feasible interior optimum point: $t_{\text{HA}} < \hat{t} < t_{\text{LA}}$. This implies Π_{HA} is strictly decreasing on $[\hat{t}, \bar{v}]$ and Π_{LA} is strictly increasing on $[\underline{v}, \hat{t}]$. Since profit is continuous at $\hat{t}$, the optimality follows. The remaining two cases can be established analogously. Finally, observe t_{HA} and t_{LA} are strictly decreasing in c, and so by construction, t_{subs}

is continuous and weakly decreasing in c.

Double-crossing comparison. We compare t_{subs} with t_{eff} in each region.

(i) $c < c_{\text{HA}}$: Given $t_{\text{subs}} = t_{\text{HA}}$, applying the definition of t_{eff} to (20) shows $\Gamma_{\text{HA}}(t) < 0$ for all $t \geq t_{\text{eff}}$. Thus, $t_{\text{subs}} < t_{\text{eff}}$ in this region of c.

(ii) $c_{\text{HA}} \leq c \leq c_{\text{LA}}$: $t_{\text{subs}} = \hat{t}$ is independent of c, whereas t_{eff} is strictly decreasing in c. Define $\underline{c}$ by $t_{\text{eff}}(\underline{c}) = \hat{t}_{\text{subs}}$, then

$$t_{\text{subs}} > t_{\text{eff}} \Leftrightarrow c > \underline{c}.$$

To verify $\underline{c} \leq c_{\text{LA}}$, we evaluate $\Gamma_{\text{LA}}(\hat{t})$ at $c = \underline{c}$, where $\underline{c} = (1-\mu)(1-F(\hat{t}))\,\text{MRL}(\hat{t})$ and $s = (1-\mu F(\hat{t}))\,\text{MRL}(\hat{t})$. Substituting into (21) and simplifying gives

$$\Gamma_{\text{LA}}(\hat{t}_{\text{subs}})\big|_{c=\underline{c}} = (1-\mu)^2\,\text{MRL}(\hat{t})\left[1-\varphi(\hat{t})\,\text{MRL}(\hat{t})\right] \geq 0,$$

where the inequality uses decreasing MRL (implies $\varphi(\hat{t})\,\text{MRL}(\hat{t}) \leq 1$). Since $\Gamma_{\text{LA}}(\hat{t})$ is decreasing in c and equals zero at $c = c_{\text{LA}}$, we conclude $\underline{c} \leq c_{\text{LA}}$.

(iii) $c > c_{\text{LA}}$: Given $t_{\text{subs}} = t_{\text{LA}}$, we continue from (21) and define $\psi(c) \equiv \Gamma_{\text{LA}}\big(t_{\text{eff}}(c),\, c\big)$. We first observe $\psi(c_{\text{LA}}) > 0$ by continuity and part (ii) above; whereas when $c \to (1-\mu)s/\mu > c_{\text{LA}}$, we have $\psi(c) < 0$ by inspecting (21). Moreover, ψ is strictly decreasing. To see this, note since $c \mapsto t_{\text{eff}}(c)$ is a strictly decreasing bijection, it suffices to show that ψ, viewed as a function of $z \equiv t_{\text{eff}}$, is increasing. Using the definition of t_{eff} in (3) to substitute away c and the reservation identity to simplify,

$$\psi(z) = s(1-\mu) - (1-\mu)\,\text{MRL}(z)\big[\mu\big(1-F(z)\big) + (1-\mu)\,\varphi(r)\,\text{MRL}(r)\big]\,,$$

where $r = r(z,0)$ is increasing in z. Observe $\text{MRL}(z)$ is decreasing, $1-F(z)$ is strictly decreasing, and $\frac{d}{dr}\big[\varphi(r)\,\text{MRL}(r)\big] = \text{MRL}''(r) \leq 0$ by the concave MRL assumption (using the standard identity $\varphi(r)\,\text{MRL}(r) = \text{MRL}'(r) + 1$). Hence, $\psi(z)$ is strictly increasing in z, as required. We conclude there exists a crossing point such that

$$t_{\text{subs}} < t_{\text{eff}} \Leftrightarrow c > \bar{c}.$$

Combining (i)-(iii) proves the proposition, with the cutoffs identified above satisfying $0 < c_{\text{HA}} < \underline{c} \leq c_{\text{LA}} < \bar{c}$.

□

Proof. **(Proposition 4)**. For part (i), the envelope theorem argument in the text has shown

$$\frac{d}{dv_0}\Pi^*_{\text{subs}} = -1 < \frac{d}{dv_0}\Pi^*_{\text{usage}}.$$

At boundary $v_0 = r_{\text{eff}}$, $\Pi^*_{\text{subs}} = \max_{t\in[\underline{v},\bar{v}]}\text{TW}(t,0) - v_0 < r_{\text{eff}} - v_0 = 0 = \Pi^*_{\text{usage}}$. Meanwhile, at $v_0 = \underline{v}$, if $\Pi^*_{\text{subs}} \geq \Pi^*_{\text{usage}}$, then the existence and uniqueness of cutoff $v_{\text{profit}} \in [\underline{v}, r_{\text{eff}})$ follow from the intermediate value theorem. Otherwise if $\Pi^*_{\text{subs}} < \Pi^*_{\text{usage}}$ for all v_0, then we set $v_{\text{profit}} = \underline{v}$. For part (ii), suppose c is sufficiently high such that $v_0 = r_{\text{eff}}$, then the same argument as above shows

$\Pi^*_{\text{subs}} < 0 = \Pi^*_{\text{usage}}$; this argument holds even when $\mu c \geq (1-\mu)s$. Meanwhile, at $c = 0$ the profit expressions (4) and (5) reduce to

$$\Pi_{\text{usage}}(t) = \frac{\Phi(u_{\text{usage}}|t) - s}{1 - G(u_{\text{usage}}|t)} = \int_{u_{\text{usage}}}^{r(t,0)} \frac{1 - G(v_i|t)}{1 - G(u_{\text{usage}}|t)}\, dv_i,$$
$$\Pi_{\text{subs}}(t) = r(t,0) - v_0,$$

where we used $s = \Phi(r(t,0)|t)$. Optimality entails $t_{\text{eff}} = t_{\text{usage}} = t_{\text{subs}} = \bar{v}$, so that $c = 0$ and Proposition 2 imply $r(\bar{v}, 0) = r_{\text{eff}} > u_{\text{usage}} \geq v_0$. Given that $1 - G(v_i|\bar{v}) < 1 - G(u_{\text{usage}}|\bar{v})$ for all $v_i \in (u_{\text{usage}}, r(\bar{v}, 0))$, we conclude $\Pi^*_{\text{subs}} > \Pi^*_{\text{usage}}$ for all sufficiently small c by continuity. □

Proof. **(Proposition 5)**. Recall $r = r(t,0)$ and $q = G(r|t)$, both being independent of n. To prove part (i), we treat n as a continuous variable. Differentiating the profit expression gives

$$\frac{\partial \Pi_{\text{cap}}}{\partial n}(t,n) = \underbrace{-\int_{v_0}^{r} G(v_i|t)^n \ln G(v_i|t)\, dv_i}_{\text{marginal WTP gain} > 0} + \underbrace{\frac{c\, q^n \ln q}{(1-q)(1-F(t))}}_{\text{marginal cost} < 0}, \tag{23}$$

where the signs follow from $G(v_i|t) \leq q < 1$ on $[v_0, r]$. When $c \to 0$, (23) implies $\partial \Pi_{\text{cap}}/\partial n > 0$ at every n, and so $n_{\text{cap}} \to \infty$. To identify the condition for $n_{\text{cap}} = 1$, we start by noticing that whenever $n_{\text{cap}} = 1$, the subscription-with-cap profit becomes

$$\Pi_{\text{cap}}(t,1) = \int_{v_0}^{r(t,0)} [1 - G(v|t)]\, dv \; - \; \frac{c}{1 - F(t)} = \Phi(v_0|t) - s - \frac{c}{1 - F(t)}.$$

where the last equality uses the reservation value identities. Its derivative is

$$\frac{\partial}{\partial t}\Pi_{\text{cap}}(t,1) = \begin{cases} \dfrac{f(t)}{(1-F(t))^2}[(1-\mu)s_0 - c] > 0 & \text{if } t \leq v_0, \\ \dfrac{f(t)}{(1-F(t))^2}[(1-\mu)\displaystyle\int_t^{\bar{v}} (1 - F(v_i))\, dv_i - c] & \text{if } t > v_0 \end{cases}$$

where the inequality follows from Assumption A: $t_{\text{eff}} > r_{\text{eff}} > v_0$, together with the t_{eff} FOC (3) and the direct-search identity $\int_{v_0}^{\bar{v}} (1 - F(v))\, dv = s_0$, gives $c/(1-\mu) < s_0$. Therefore the optimum lies in $t > v_0$, where the FOC coincides with that of t_{eff}, giving $t_{\text{cap}} = t_{\text{eff}}$. Since $t_{\text{eff}} > r_{\text{eff}} > v_0$ under Assumption A, the outcome is in the high-acceptance regime, so $G(v_i|t_{\text{eff}}) = \mu F(v_i)$ on $[v_0, r]$. Substituting into (23), the consistency requirement for $n_{\text{cap}} = 1$ to arise in equilibrium is $\frac{\partial \Pi_{\text{cap}}}{\partial n}(t_{\text{eff}}, 1) \leq 0$, or equivalently,

$$c \geq (1 - \mu F(r))(1 - F(t_{\text{eff}})) \int_{v_0}^{r} \frac{F(v_i)}{F(r)} \cdot \frac{\ln(\mu F(v_i))}{\ln(\mu F(r))}\, dv_i, \tag{24}$$

which requires c to be sufficiently large relative to μ. Finally, part (ii) follows from the same argument as Proposition 4, with the observation that: (a) as $c \to 0$, $\Pi^*_{\text{cap}} \to \Pi^*_{\text{subs}}$ since the cap is unconstrained, so $\Pi^*_{\text{cap}} > \Pi^*_{\text{usage}}$ for c small enough; (b) at $v_0 = r_{\text{eff}}$, $\Pi^*_{\text{cap}} < 0 \leq \Pi^*_{\text{usage}}$, and so $\Pi^*_{\text{cap}} < \Pi^*_{\text{usage}}$ for v_0 large enough. □

Proof. **(Proposition 6)**. We start by characterizing the equilibrium profits before proving Propo-

sition 6. By Corollary 2, the analysis of the Stage-1 competition subgame (for given business models) proceeds like a standard Hotelling model with providers competing in offering utilities, facing per-consumer margin $\pi_{\text{usage}}(t_{\text{usage}}, u_k)$ or $\pi_{\text{subs}}(t_{\text{subs}}, u_k)$. For notational simplicity, denote

$$x^\circ = \tfrac{1}{2} + \frac{u_k - u_{-k}}{2\tau}.$$

In any equilibrium, $u_k \leq r_{\text{eff}}$, the maximized total welfare, as otherwise the profit is strictly negative.

Symmetric equilibria. When both providers adopt the same business model ω, each provider k chooses u_k to maximize profit

$$\pi_\omega(t_\omega, u_k)\, H(x^\circ).$$

Denote the equilibrium utility level offered in each regime as u_ω for $\omega \in \{\text{subs}, \text{usage}\}$. The corresponding first-order conditions, after imposing symmetry, are

$$\frac{1}{2}\frac{\partial \pi_\omega}{\partial u_k}(t_\omega, u_\omega) + \pi_\omega(t_\omega, u_\omega)\,\frac{h(1/2)}{2\tau} = 0. \tag{25}$$

Here, by the same calculation as (6), the equilibrium per-consumer margins satisfy

$$\begin{aligned}
\frac{\partial \pi_{\text{subs}}}{\partial u_k}(t_{\text{subs}}, u_{\text{subs}}) &= -1 \\
\frac{\partial \pi_{\text{usage}}}{\partial u_k}(t_{\text{usage}}, u_{\text{usage}}) &= -1 + \frac{\mu f(u_{\text{usage}})}{1 - \mu F(u_{\text{usage}})}\,\pi_{\text{usage}}(t_{\text{usage}}, u_{\text{usage}}) > -1,
\end{aligned}$$

where the usage derivative uses $G(u_k|t_{\text{usage}}) = \mu F(u_k)$, valid in the high-acceptance regime (recall Assumption A and Proposition 2 imply $t_{\text{usage}} = t_{\text{eff}} > r_{\text{eff}} \geq u_k$). Substituting the derivatives into (25), we conclude the equilibrium per-consumer margin satisfies $\pi_{\text{subs}}(t_{\text{subs}}, u_{\text{subs}}) > \pi_{\text{usage}}(t_{\text{usage}}, u_{\text{usage}})$. Therefore,

$$\Pi^*(\text{subs}; \text{subs}) = \frac{1}{2}\pi_{\text{subs}}(t_{\text{subs}}, u_{\text{subs}}) = \frac{\tau}{2\,h(1/2)} > \Pi^*(\text{usage}; \text{usage}) > 0 \tag{26}$$

Asymmetric equilibrium. Without loss of generality we let $\omega_1 = \text{usage}$ and $\omega_2 = \text{subs}$. Denote $\pi_1(u_1) = \pi_{\text{usage}}(t_{\text{usage}}, u_1)$, and

$$w_{\text{subs}} = \pi_{\text{subs}}(t_{\text{subs}}, u_2) + u_2 = r(t_{\text{subs}}, 0) - \frac{c}{1 - F(t_{\text{subs}})} \cdot \frac{1}{1 - G(r(t_{\text{subs}}, 0)|t_{\text{subs}})}.$$

Note $w_{\text{subs}} < r_{\text{eff}}$. Then, the first-order conditions for the equilibrium utility offers are

$$\begin{aligned}
\frac{d}{du_1}\left[\pi_1(u_1) H(x^\circ)\right] &= 0 \implies \left(-1 + \frac{\mu f(u_1)\pi_1(u_1)}{1 - \mu F(u_1)}\right)\frac{H(x^\circ)}{h(x^\circ)} + \frac{\pi_1(u_1)}{2\tau} = 0 \qquad (27) \\
\frac{d}{du_2}\left[(w_{\text{subs}} - u_2)(1 - H(x^\circ))\right] &= 0 \implies -\frac{1 - H(x^\circ)}{h(x^\circ)} + \frac{w_{\text{subs}} - u_2}{2\tau} = 0.
\end{aligned}$$

Combining the two equations with $x^\circ = \frac{1}{2} + \frac{u_1 - u_2}{2\tau}$ yields the equilibrium outcome. We want to identify the condition on τ for $x^\circ = 1$ to occur in equilibrium. For that to hold, we must have provider 2 earning zero margin, so $u_2 = w_{\text{subs}}$. Therefore, we just need the solution to (27) to

satisfy $u_1 \geq w_{\text{subs}} + \tau$, i.e.,

$$\left[\left(-1 + \frac{\mu f(u_1)\pi_1(u_1)}{1 - \mu F(u_1)}\right)\frac{1}{h(1)} + \frac{\pi_1(u_1)}{2\tau}\right]_{u_1 = w_{\text{subs}} + \tau} > 0.$$

If $\tau \to 0$, then this inequality holds strictly given $\pi_1(w_{\text{subs}}) > \pi_1(r_{\text{eff}}) = 0$. By continuity, this strict inequality holds for all strictly positive τ close to zero. Thus, there exists a strictly positive cutoff $\bar{\tau} > 0$ such that $\tau < \bar{\tau}$ implies

$$\Pi^*(\text{usage}; \text{subs}) > 0 = \Pi^*(\text{subs}; \text{usage})$$

and $\Pi^*(\text{usage}; \text{subs})$ is bounded away from zero.

Comparison. We now compare all four configurations. For all strictly positive $\tau < \bar{\tau}$, we have

$$\Pi^*(\text{subs}; \text{usage}) = 0 < \Pi^*(\text{usage}; \text{usage}) < \Pi^*(\text{subs}; \text{subs})$$

by (26). Moreover, as $\tau \to 0$, we have $\Pi^*(\text{subs}; \text{subs}) \to 0 < \Pi^*(\text{usage}; \text{subs})$. Combining these inequalities establishes the profit comparison claimed. □

ONLINE APPENDIX

AI-delegated search: pricing, search quality, and business model implications

In this online appendix we provide additional results referred to but not included in the main paper.

A Analysis with uniform distribution

With uniform distribution, $v_0 = 1 - \sqrt{2s_0}$. We first note

$$1 - G(v) = (1-\mu)\min\left\{\frac{1-v}{1-t}, 1\right\} + \mu(1-v) = \begin{cases} 1 - \mu v, & \text{if } v \le t, \\ (1-v)\dfrac{1-\mu t}{1-t}, & \text{if } v \ge t. \end{cases}$$

Computing $\Phi(v|t)$,

$$\Phi(v|t) = \begin{cases} \dfrac{(1-\mu t)(1-v)^2}{2(1-t)}, & \text{if } v \ge t, \\ \dfrac{t(1-\mu) + \mu v^2 - 2v + 1}{2}, & \text{if } v < t. \end{cases}$$

We now solve $\Phi(r) = s + p$ for $r(t,p)$ using the quadratic formula. For the LA branch with $r \ge t$, we get

$$r = 1 - \sqrt{\frac{2(s+p)(1-t)}{1-\mu t}} \ge t \quad \iff \quad s + p \le \frac{(1-\mu t)(1-t)}{2}$$

For the HA branch with $r \le t$, we get

$$r = \frac{1}{\mu}\left(1 - \sqrt{(1-\mu)(1-\mu t) + 2\mu(s+p)}\,\right) \le t \quad \iff \quad s + p \ge \frac{(1-\mu t)(1-t)}{2}$$

The boundary $s + p = \frac{(1-\mu t)(1-t)}{2}$ is exactly where the two expressions coincide (i.e. where $r = t$), so $r(t,p)$ is continuous.

Usage pricing. Profit as a function of threshold t and target utility $u \in [v_0, r_{\text{eff}}]$ is

$$\Pi_{\text{usage}}(t,u) = \begin{cases} \frac{1-u}{2} - \frac{s(1-t)+c}{1-\mu t}, & \text{if } t \le u \\ \frac{1}{1-\mu u}\left(\frac{t(1-\mu)+\mu u^2 - 2u + 1}{2} - \frac{c}{1-t}\right), & \text{if } t > u. \end{cases}$$

with optimal threshold being $t_{\text{eff}} = 1 - \sqrt{\frac{2c}{1-\mu}} > r_{\text{eff}} > u$ by Proposition 2. Then, the FOC that pins down the optimal target utility u_{usage} is

$$-\mu u^2 + 2u - \frac{2}{\mu} + 1 + t_{\text{eff}}(1-\mu) - \frac{2c}{1-t_{\text{eff}}} = 0. \tag{28}$$

Observe if $\mu \to 0$ then the FOC has no solution, in which case $u_{\text{usage}} = v_0$.

Subscription pricing. Profit function after letting $T = r(t,0) - v_0$ is

$$\Pi_{\text{subs}}(t) = \begin{cases} 1 - v_0 - \frac{2s(1-t)+c}{\sqrt{2s(1-t)(1-\mu t)}}, & \text{if } t \le \hat{t}_{\text{subs}}, \\ \frac{1-\sqrt{(1-\mu)(1-\mu t)+2\mu s}}{\mu} - v_0 - \frac{c}{(1-t)\sqrt{(1-\mu)(1-\mu t)+2\mu s}}, & \text{if } t > \hat{t}_{\text{subs}}. \end{cases}$$

where the boundary is given by

$$\frac{1}{2}(1-\mu\hat{t}_{\text{subs}})(1-\hat{t}_{\text{subs}}) = s \implies \hat{t}_{\text{subs}} = \frac{(1+\mu) - \sqrt{(1-\mu)^2 + 8\mu s}}{2\mu}.$$

The optimal threshold is

$$t_{\text{subs}} = \begin{cases} t_{\text{HA}} & \text{if } c < c_{\text{HA}} \quad \text{(Branch HA interior)}, \\ \hat{t}_{\text{subs}} & \text{if } c_{\text{HA}} \le c \le c_{\text{LA}} \quad \text{(boundary)}, \\ t_{\text{LA}} & \text{if } c > c_{\text{LA}} \quad \text{(Branch LA interior)}. \end{cases}$$

where t_{LA} and t_{HA} are respectively the interior solutions to the LA and HA branches of $\Pi_{\text{subs}}(t)$:

$$c_{\text{LA}} = \frac{4s^2(1-\mu)}{(1-\mu)(1-\mu\hat{t}_{\text{subs}}) + 4\mu s}, \qquad c_{\text{HA}} = \frac{4s^2(1-\mu)}{3\mu(1-\hat{t}_{\text{subs}})(1-\mu) + 2\big[(1-\mu)^2 + 2\mu s\big]},$$

When $\mu \to 0$, the solution admits a simple closed-form:

$$t_{\text{subs}} = \begin{cases} t_{\text{eff}} & \text{if } c < 2s^2, \\ 1 - 2s & \text{if } 2s^2 \le c \le 4s^2, \\ 1 - \frac{c}{2s} & \text{if } c > 4s^2. \end{cases}$$

We now prove the cutoff result on profit comparison claimed in Section 4. Formally:

Lemma 5. *Suppose $\mu > 0$ is sufficiently small and $F \sim U[0,1]$. Then there exists a cost threshold $c_{\text{profit}} > 0$ such that*

$$\Pi^*_{\text{usage}} < \Pi^*_{\text{subs}} \Leftrightarrow c < c_{\text{profit}}.$$

The proof distinguishes two cases: $c_{\text{profit}} \in (0, 2s^2]$ when $s \ge 1/8$, and $c_{\text{profit}} \in (2s^2, 4s^2)$ when $s < 1/8$.

The equilibrium profits extend continuously to the $\mu \to 0$ limit, so we expand around it to obtain the result for small $\mu > 0$. To prove the lemma, we focus on $s_0 = s$ ($s_0 > s$ does not change the analysis and just expands the region for $\Pi^*_{\text{usage}} < \Pi^*_{\text{subs}}$). We note from (28) that when $\mu \to 0$ then $u_{\text{usage}} = v_0$.

By a first-order Taylor expansion locally near $\mu = 0$:

$$\begin{aligned} \Pi^*_{\text{usage}} &= \Pi^*_{\text{usage}}|_{\mu=0} + \frac{d}{d\mu}\Pi^*_{\text{usage}}|_{\mu=0} \times \mu + O(\mu^2) \\ &= \sqrt{2s} - s - \sqrt{2c} + \left[(2\sqrt{s} - \frac{1}{\sqrt{2}})\sqrt{c} + \left(\sqrt{2s} - 2\right) s\right]\mu + O(\mu^2). \end{aligned} \tag{29}$$

and

$$\Pi^*_{\text{subs}} = \begin{cases} \sqrt{2s} - s - \sqrt{2c} + \left[(2s-1)\frac{\sqrt{c}}{\sqrt{2}} + \frac{3c}{4} + \frac{s^2}{2} - s\right]\mu + O(\mu^2) & \text{if } c \le 2s^2 \\ \sqrt{2s} - 2s - \frac{c}{2s} + 2s(2s-1)\mu + O(\mu^2) & \text{if } c \in [2s^2, 4s^2] \\ \sqrt{2s} - 2\sqrt{c} + \sqrt{c}\,(\frac{c}{2s} - 1)\mu + O(\mu^2) & \text{if } c \ge 4s^2 \end{cases}, \tag{30}$$

which is continuous in c. Then, collecting the intercept and slope term, we denote the profit difference as

$$\Pi^*_{\text{usage}} - \Pi^*_{\text{subs}} = Z(\mu, c) + O(\mu^2)$$

where $Z(\mu, c)$ is first-order Taylor approximation and it is continuous in (μ, c) from (29) and (30).

Then, the crossing point c° is defined as the solution to

$$Z(\mu, c^\circ) = 0$$

and its existence $c^\circ \in (0, 4s^2)$ is guaranteed by the intermediate value theorem because $Z(\mu, 0) < 0$; whereas for all $c \ge 4s^2$, $\lim_{\mu \to 0} Z(\mu, c) = (2 - \sqrt{2})$ and so $Z(\mu, c) > 0$ for sufficiently small μ.

It remains to establish uniqueness of c°. Recall

$$Z(\mu, c) = \begin{cases} \left[2\sqrt{cs} + \sqrt{2}\, s^{3/2} - s - s\sqrt{2c} - \frac{3c}{4} - \frac{s^2}{2}\right]\mu & \text{for } c \in (0, 2s^2] \\ s - \sqrt{2c} + \frac{c}{2s} + \left[(2\sqrt{s} - \frac{1}{\sqrt{2}})\sqrt{c} + \sqrt{2}\, s^{3/2} - 4s^2\right]\mu & \text{for } c \in (2s^2, 4s^2) \end{cases}.$$

Then

$$\text{for } c \in (0, 2s^2]\text{: } \frac{\partial}{\partial c} Z(\mu, c) = \left[\frac{\sqrt{s}}{\sqrt{c}}\left(1 - \sqrt{\tfrac{s}{2}}\right) - \frac{3}{4}\right]\mu > \left[\frac{\sqrt{2s}}{\sqrt{2s} - s}\left(1 - \sqrt{\tfrac{s}{2}}\right) - \frac{3}{4}\right]\mu = \frac{\mu}{4} > 0$$

where the inequality used Assumption A; whereas

$$\text{for } c \in (2s^2, 4s^2)\text{: } \quad \frac{\partial}{\partial c} Z(\mu, c) = \left[\frac{1}{2s} - \frac{1}{\sqrt{2c}}\right] + \left[\sqrt{s} - \frac{1}{2\sqrt{2}}\right]\frac{\mu}{\sqrt{c}} > 0 \quad \text{if } s \ge \frac{1}{8}$$

To proceed, we note at the boundary, we have

$$Z\left(\mu, 2s^2\right) = [3\sqrt{2s} - 1 - 4s]s\mu \ge 0 \iff s \ge \frac{1}{8}.$$

Then, the following are immediate:

1) If $s \ge 1/8$, then for all $c \in (0, 4s^2)$ the derivative is $\frac{\partial}{\partial c} Z(\mu, c) > 0$. By the intermediate value theorem, there exists a unique crossing point $c^\circ \in (0, 2s^2]$.

2) If $s < 1/8$, then for $c \in (0, 2s^2]$, $Z(\mu, c) < 0$; whereas for $c \in (2s^2, 4s^2)$, the second derivative is

$$\frac{\partial^2}{\partial c^2} Z(\mu, c) = \frac{1}{(2c)^{3/2}}\left[1 + (\frac{1}{2\sqrt{2}} - \sqrt{s})\mu\right] > 0.$$

By convexity, there exists a unique crossing point $c^\circ \in (2s^2, 4s^2)$.

B Alternative formulation of subscription-with-cap

This appendix revisits the subscription-with-cap model in Section 4.2 by assuming instead that the cap is imposed in the form of a cap N on the total number of inner searches.

Recall that we assume no premature-stopping: if the inner-search cap N is reached while the agent is in the middle of a query's inner search, the agent completes that query (it keeps drawing until some $\hat{v}_i \geq t$ and delivers that recommendation) before the service terminates. The alternative (stopping mid-search and handing back the best sub-threshold option seen) would make the *last* query's value a draw from a different distribution than $G(\cdot \,|\, t)$, breaking the i.i.d. property.

Consumer problem

Two facts from the baseline are used throughout. First, the per-query inner-search count is geometric. Because the unconditional signal distribution is F, each inner draw clears the threshold independently with probability $1 - F(t)$. Hence the number of inner searches the agent performs to deliver one query,

$$K \;\sim\; \mathrm{Geom}\big(1 - F(t)\big), \qquad \Pr(K = k) = F(t)^{\,k-1}\big(1 - F(t)\big), \quad \mathbb{E}[K] = \frac{1}{1 - F(t)}. \tag{31}$$

Distinct queries draw fresh, independent options, so the per-query counts $K_1, K_2, \ldots$ are i.i.d.

Second, as a consequence of no premature-stopping, the values of recommended options are independent of the inner-search count. Conditional on $\hat{v}_i \geq t$, the posterior CDF of the recommended option's match value is

$$G(v \,|\, t) = \mu F(v) + (1 - \mu)\, \frac{F(v) - F(t)}{1 - F(t)}\, \mathbf{1}_{\{v \geq t\}},$$

and the realization of the values are i.i.d. With perfect recall and i.i.d. draws, the consumer's optimal rule is a cutoff at the reservation value $r \equiv r(t, 0)$ (Weitzman, 1979), as in Section 4.2.

The following result derives the consumer willingness to pay and the expected number of queries for given threshold t and inner-search cap N.

Lemma 6. *A subscription with inner-search cap N induces a random query cap M that satisfies*

$$M - 1 \;\sim\; \mathrm{Binom}\big(N - 1,\, 1 - F(t)\big) \tag{32}$$

and M is independent of the realization of match values in the consumer's query sequence. Consequently the consumer willingness to pay and expected query volume under the inner-search cap N equal their query-cap counterparts averaged over the induced random query cap M:

$$r - v_0 - \mathbb{E}\Big[\int_{v_0}^{r} G(v \,|\, t)^{\,M}\, dv\Big] = r - v_0 - \int_{v_0}^{r} G(v \,|\, t)\, \big(F(t) + (1 - F(t))\, G(v \,|\, t)\big)^{N-1}\, dv, \tag{33}$$

and

$$\alpha = \mathbb{E}\Big[\frac{1 - q^M}{1 - q}\Big] = \frac{1 - q\,(F(t) + (1 - F(t))q)^{N-1}}{1 - q}. \tag{34}$$

Proof. Represent the inner-search stream as a sequence of i.i.d. Bernoulli trials, one per inner

draw, each a "success" (clears t, completing a query) with probability $1 - F(t)$ by (31). Query m completes at the trial of the m-th success, and the total number of inner searches up until query m is $S_m = K_1 + \cdots + K_m$. Given no premature-stopping, the AI agent runs until the first success at or after trial N, so the number of completed queries, in the cap-binding event, is

$$M = \min\{m : S_m \geq N\}.$$

Equivalently, $M - 1$ equals the number of queries that complete strictly before the cap is reached, i.e. the number of successes among the first $N - 1$ trials, which is $\text{Binom}(N - 1, 1 - F(t))$; this gives (32). Independence holds because M is a function of the inner-search counts $\{K_j\}$ only, which are independent of the realized match values. The random horizon M satisfies

$$\mathbb{E}\big[z^M\big] = z\,\big(F(t) + (1 - F(t))\,z\big)^{N-1}, \qquad \mathbb{E}[M] = 1 + (N-1)\big(1 - F(t)\big). \tag{35}$$

Recall that the consumer applies the cutoff r to i.i.d. $G(\cdot\,|\,t)$ draws. Therefore, conditional on a realized horizon M, her WTP and expected number of queries are exactly those under fixed query cap M. Averaging over M gives (33) and (34). □

Provider's problem

We now solve the AI provider's problem under a subscription with an inner-search cap and establish the equilibrium analogue of Proposition 5. We keep Assumption A.

With identical consumers the provider extracts the full willingness to pay (33) through the lump-sum fee. Total compute is $c\,\mathbb{E}[K_1 + \cdots + K_Q]$, where $Q = \min(\theta, M)$ is the number of delivered queries, with θ the first accepted query and M the cap horizon (32). No premature-stopping implies independence, so $\mathbb{E}[K_j\,\mathbf{1}\{Q \geq j\}] = \mathbb{E}[K_j]\,\Pr(Q \geq j)$. Summing over j gives $\mathbb{E}[K_j]\,\mathbb{E}[Q] = \alpha/(1 - F(t))$. The provider therefore chooses t and N to maximize profit

$$\Pi_{\text{icap}}(t, N) = r - v_0 - \int_{v_0}^{r} G(v\,|\,t)\big(F(t) + (1 - F(t))G(v\,|\,t)\big)^{N-1}\,dv - \frac{c\,\alpha}{1 - F(t)}. \tag{36}$$

The following proposition reproduces Proposition 5 in the inner-search-cap environment, confirming the main paper's footnote claim that the result continues to hold when the provider caps total inner searches rather than queries.

Proposition OA.1 (Subscription with an inner-search cap)**.** *Under subscription pricing with an inner-search cap $N \in \{1, 2, \ldots\}$, the equilibrium satisfies the following properties:*

(i) *If $c \to 0$, then $N_{\text{icap}} \to \infty$ and $t_{\text{icap}} \to t_{\text{subs}}$. If c is large enough, then $N_{\text{icap}} = 1$ and $t_{\text{icap}} = t_{\text{eff}}$.*

(ii) *If v_0 is large enough, $\Pi^*_{\text{icap}} < \Pi^*_{\text{usage}}$; if c is small enough, then $\Pi^*_{\text{icap}} > \Pi^*_{\text{usage}}$.*

Proof. Recall r and q are independent of N. To prove part (i), treat N as a continuous variable.

Differentiating (36),

$$\frac{\partial \Pi_{\text{icap}}}{\partial N}(t,N) = \underbrace{-\int_{v_0}^{r} G(v\,|\,t)\big(F(t)+(1-F(t))G(v\,|\,t)\big)^{N-1}\ln\big(F(t)+(1-F(t))G(v\,|\,t)\big)\,dv}_{\text{marginal WTP gain}>0} + \underbrace{\frac{c\,q\big(F(t)+(1-F(t))q\big)^{N-1}\ln\big(F(t)+(1-F(t))q\big)}{(1-q)(1-F(t))}}_{\text{marginal compute}<0}, \tag{37}$$

where the signs follow from $F(t)+(1-F(t))G(v\,|\,t) \le F(t)+(1-F(t))q < 1$ on $[v_0, r]$. When $c \to 0$, the marginal compute vanishes while the WTP gain stays positive, so (37) is positive at every N and hence $N_{\text{icap}} \to \infty$; the cap becomes slack and $t_{\text{icap}} \to \arg\max_t \Pi_{\text{subs}}(t) = t_{\text{subs}}$.

To identify the condition for $N_{\text{icap}} = 1$, note that whenever $N_{\text{icap}} = 1$ the cap delivers one query ($\alpha = 1$), so

$$\Pi_{\text{icap}}(t,1) = \int_{v_0}^{r(t,0)} [1-G(v|t)]\,dv \;-\; \frac{c}{1-F(t)} = \Phi(v_0|t) - s - \frac{c}{1-F(t)}.$$

This expression is exactly the same as $\Pi_{\text{cap}}(t,n)$ with $n=1$, and so Proposition 5(i) in the main text immediately applies.

For part (ii), the cap problem nests pure subscription ($N \to \infty$), so $\Pi^*_{\text{icap}} \ge \Pi^*_{\text{subs}}$; since $\Pi^*_{\text{subs}} > \Pi^*_{\text{usage}}$ for small c (Proposition 4 of the main paper), $\Pi^*_{\text{icap}} > \Pi^*_{\text{usage}}$ for c small enough. For the outside-option claim, we again note at $v_0 = r_{\text{eff}}$, $\Pi^*_{\text{icap}} < 0 \le \Pi^*_{\text{usage}}$, and so $\Pi^*_{\text{icap}} < \Pi^*_{\text{usage}}$ for v_0 large enough. □

C Details of extensions

C.1 Details of learning and decreasing AI signal noise

Consumer's problem

Throughout, fix the inner-search threshold t and per-query fee p (with $p = 0$ under subscription).

As preliminaries, we first let $H(v|t) \equiv \frac{F(v)-F(t)}{1-F(t)}\mathbf{1}_{\{v\geq t\}}$ be the law of an informatively recommended value, and decompose

$$G_m(v|t) = \mu_m F(v) + (1-\mu_m)H(v|t), \qquad \Phi_m(v|t) = \Phi_F(v) + (1-\mu_m)\,\Delta(v|t),$$

where $\Phi_F(v) \equiv \int_v^{\bar{v}}(1-F(v_i))\,dv_i$ and $\Delta(v|t) \equiv \int_v^{\bar{v}}[F(v_i) - H(v_i|t)]\,dv_i$. Since $F \geq H$, $\Delta(\cdot|t)$ is positive and decreasing, and it is strictly positive on $[\underline{v}, \bar{v})$ when $t > \underline{v}$. Because μ_m is decreasing and $F \geq H$, the posterior improves in the first-order stochastic dominance sense as m grows, which implies for all $m' \geq m$:

$$G_{m'}(\cdot|t) \leq G_m(\cdot|t) \qquad \text{and} \qquad \Phi_{m'}(\cdot|t) \geq \Phi_m(\cdot|t).$$

The following analysis completes the proof of Lemma 3.

Value function and reservation property. A consumer at the beginning of step m holding best-so-far value w faces a simple choice: accept, ending the search with payoff w, or pay $s+p$ to issue the m-th query and continue optimally. Writing $V_m(w)$ as the value function, i.e., her resulting optimal expected payoff. The Bellman equation is

$$V_m(w) = \max\{w,\, C_m(w)\}, \qquad C_m(w) \equiv -(s+p) + \mathbb{E}_{G_m}[V_{m+1}(\max\{w, v_i\})]\,, \tag{38}$$

where $C_m(w)$ is the *continuation value*: the expected payoff from paying the per-query cost $s+p$, drawing the m-th recommendation $v_i \sim G_m(\cdot|t)$, and continuing optimally from step $m+1$. The value function is bounded

$$w \leq V_m(w) \leq \max\{w, \bar{v} - (s+p)\}.$$

The induced optimal rule is to accept at the first step with $V_m(w) = w$.[1]

Each V_m is increasing and 1-Lipschitz, that is,

$$0 \leq V_m(w') - V_m(w) \leq w' - w \qquad \text{for all } w' \geq w \text{ and } w, w' \in [\underline{v}, \bar{v}] \tag{39}$$

Hence, for $\underline{v} \leq w < w' < \bar{v}$, using (38) and (39), we get

$$\begin{aligned} C_m(w') - C_m(w) &= \int_{\underline{v}}^{\bar{v}} [V_{m+1}(\max\{w', v_i\}) - V_{m+1}(\max\{w, v_i\})]\; dG_m(v_i) \\ &\leq G_m(w'|t)\,(w' - w) < w' - w, \end{aligned}$$

so $w - C_m(w)$ is strictly increasing on $[\underline{v}, \bar{v})$. At the upper end $\bar{v} - C_m(\bar{v}) = s + p > 0$; therefore

[1] The one-step rule is optimal in the general optimal-stopping problem as the environment is monotone. See, e.g., Chow, Robbins, and Siegmund, *Great Expectations: The Theory of Optimal Stopping* (1971), ch. 3.

the reservation value $r_m \in (\underline{v}, \bar{v})$ is unique and pinned down by

$$r_m = C_m(r_m)$$

if the lower-end boundary condition is satisfied (i.e., $C_m(\underline{v}) - \underline{v} \geq \Phi_m(\underline{v}|t) - (s+p) > 0$); otherwise we let $r_m = \underline{v}$ without loss of generality. By construction,

$$w > r_m \Leftrightarrow w > C_m(w).$$

This proves the reservation property stated in Lemma 3.

Monotonicity in m and recursion Fix arbitrary $m \geq 1$. For all $w > r_{m+1}$, we have $V_{m+1}(w) = w > C_{m+1}(w)$ by definition, and so

$$\begin{aligned}
w &> -s - p + \mathbb{E}_{G_{m+1}}[V_{m+2}(\max\{w, v_i\})] \\
&\geq -s - p + \mathbb{E}_{G_{m+1}}[\max\{w, v_i\}] && \text{(definition of } V_{m+2}) \\
&\geq -s - p + \mathbb{E}_{G_m}[\max\{w, v_i\}] && \text{(first-order stochastic dominance)} \\
&= -s - p + \mathbb{E}_{G_m}[V_{m+1}(\max\{w, v_i\})] && \text{(given } \max\{w, v_i\} \geq r_{m+1}) \\
&= C_m(w)
\end{aligned}$$

Therefore, $w > C_m(w)$ for all $w > r_{m+1}$, implying the solution to $r_m = C_m(r_m)$ must satisfy $r_m \leq r_{m+1}$. This proves that the sequence of step-dependent reservation values $\{r_m\}_{m \geq 1}$ is increasing. Moreover, any r_m is bounded above by $\bar{v} - s - p$, and so the sequence converges by the monotone convergence theorem.

Closed form and the recursion. Fix arbitrary $m \geq 1$. As a preliminary step, we claim

$$r_m = C_m(\underline{v}) \qquad \text{and} \qquad V_m(\cdot) = \max\{\cdot, r_m\} \quad \text{everywhere.}$$

For the former, we note for all $w \leq r_{m+1}$, the definition of value function V_{m+1} implies

$$C_m(w) = -s - p + \mathbb{E}_{G_m}[V_{m+1}(\max\{w, v_i\})] = -s - p + \mathbb{E}_{G_m}[V_{m+1}(\max\{\underline{v}, v_i\})] = C_m(\underline{v}),$$

reflecting that any carried best-so-far value w that is below r_m will never trigger acceptance in step $m+1$. Since $r_m \leq r_{m+1}$ by monotonicity proven, we have $r_m = C_m(r_m) = C_m(\underline{v})$, as required. For the latter, any $w > r_m$ implies $V_m(w) = w$; whereas any $w \leq r_m$ implies $w \leq r_{m+1}$ by monotonicity and so

$$V_m(w) = \max\{w,\, C_m(\underline{v})\} = \max\{w, r_m\} \quad \text{for } w \leq r_m,$$

as required. Combining these properties, we get the recursion relation claimed in Lemma 3:

$$\begin{aligned}
r_m = C_m(\underline{v}) &= -(s+p) + \mathbb{E}_{G_m}[V_{m+1}(v_i)] \\
&= -(s+p) + \mathbb{E}_{G_m}[\max\{v_i, r_{m+1}\}] \\
&= -(s+p) + r_{m+1} + \Phi_m(r_{m+1}|t).
\end{aligned}$$

Provider's problem

The sign-up value is r_1 and the expected number of queries is

$$\alpha = \sum_{m \geq 1} \prod_{j=1}^{m-1} G_j(r_{j+1}|t),$$

the learning analogue of $\alpha(t,p)$ (empty product equal to one). Because each query triggers a fresh inner search, the expected compute cost per query is $\frac{c}{1-F(t)}$, independent of the query index m. The three profit maximization problems of the baseline carry over:

- *First-best (two-part tariff benchmark).* A constant per-query fee $p = \frac{c}{1-F(t)}$ equates the consumer's private cost $s+p$ with the social cost $s + \frac{c}{1-F(t)}$ at *every* query. Total welfare then equals r_1, and the planner maximizes r_1 over the single threshold t; the maximal profit is $r_1 - v_0$, attained by the two-part tariff $T = r_1 - v_0$, $p = \frac{c}{1-F(t)}$ (see Section 4).
- *Usage pricing.* $\Pi_{\text{usage}} = \big(p - \frac{c}{1-F(t)}\big)\alpha$, maximized over (t,p) subject to participation $r_1 \geq v_0$.
- *Subscription pricing.* $p = 0$, lump-sum fee $T = r_1 - v_0$, and $\Pi_{\text{subs}} = r_1 - v_0 - \frac{c}{1-F(t)}\,\alpha$, maximized over t.

Numerical solution. The numerical analysis adopts the parametric schedule $\mu_m = \mu/m^\eta$ described in Section 6.1. For the uniform example $F = U[0,1]$, the posterior $G_m(\cdot|t)$, the gain $\Phi_m(\cdot|t)$, and the reservation sequence are computed directly.

Fixing (t,p,η), the sequence $\{r_m\}$ is built by the backward recursion (9) from a distant horizon M, with terminal value r_M pinned down by the stationary non-learning counterpart:

$$\int_{r_M}^{\bar{v}} 1 - G_M(v_i|t)dv_i = p + s$$

the recursion is a contraction with modulus $F(\bar{x}) < 1$ for any $\bar{x} \in (r_\infty, \bar{v})$, so the truncation error in r_1 is of order $F(\bar{x})^M$ and is negligible for a moderately large M. The sign-up value r_1 and the expected number of queries α are then accumulated forward from the stored $\{r_m\}$.

For each (c,η), the first-best and subscription problems are one-dimensional maximizations over t (grid search refined by a local optimizer), and the usage program is a two-dimensional maximization over (t,p) on a grid masked by the participation constraint and refined locally. Figure 7 reports the resulting equilibrium thresholds and profits across learning rates.

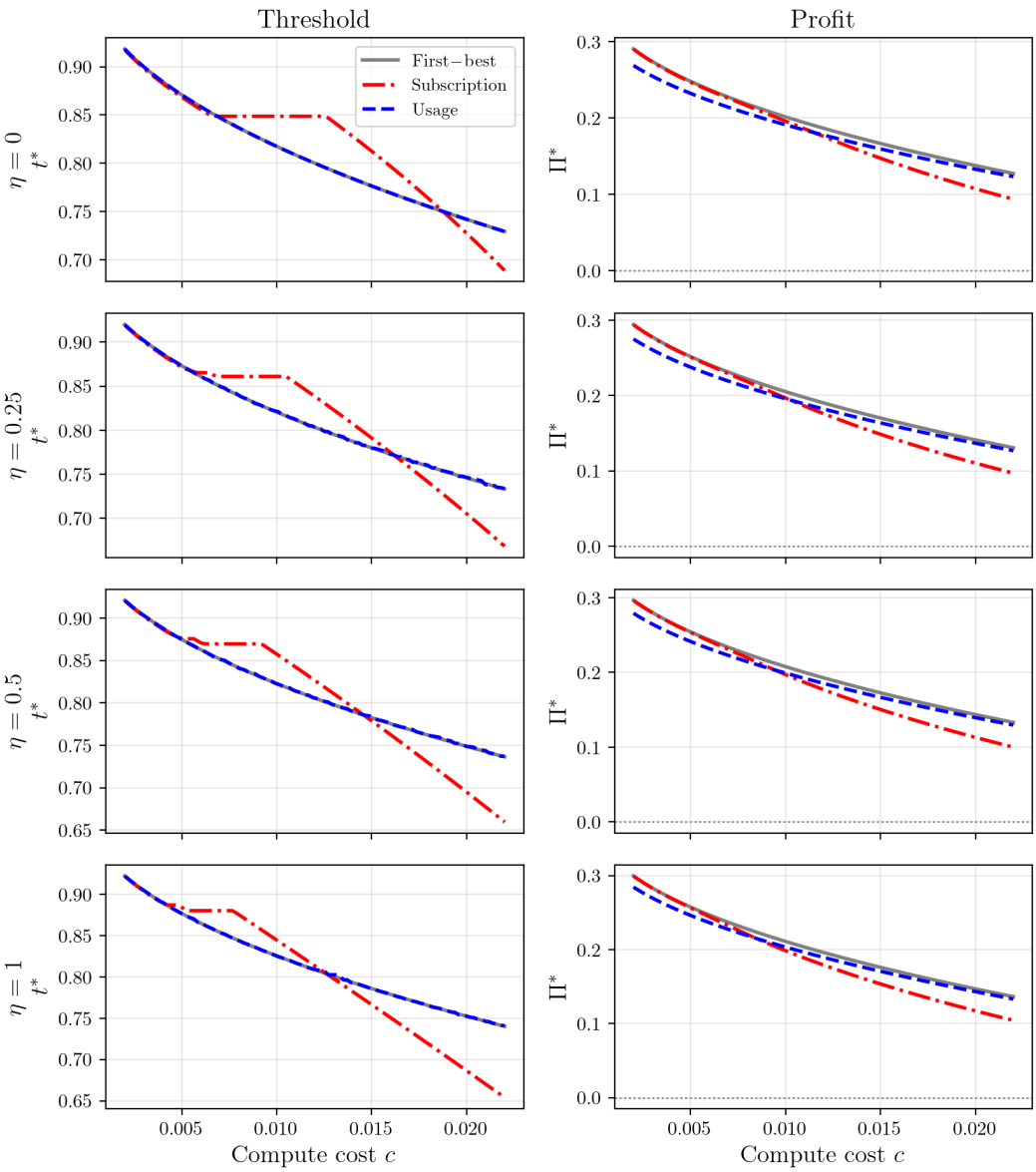

Figure 7: Equilibrium inner-search threshold t^* (left) and profit Π^* (right) versus compute cost c under the first-best two-part tariff, usage pricing, and subscription pricing, for learning rates $\eta \in \{0, 0.25, 0.5, 1\}$ (top to bottom). The top row ($\eta = 0$) reproduces the constant-noise baseline. The parameters are $F \sim U[0,1]$, $\mu = 0.4$, $s = 0.05$, and $s_0 = 0.1$.

C.2 Details of blind acceptance

As a preliminary, we recall the if-and-only-if condition for consumer blind acceptance is $B(t) - p \geq r(t,p)$. Using the identity $B(t) = \underline{v} + \Phi(\underline{v}|t)$ with

$$\Phi(\underline{v}|t) = \int_{\underline{v}}^{t} (1 - \mu F(v_i))\, dv_i + (1 - \mu F(t))\, \mathrm{MRL}(t)$$

we conclude

$$B(t) - p \geq r(t,p) \quad \Leftrightarrow \quad s \;\geq\; \int_{\underline{v}}^{r(t,p)} G(v_i|t)\, dv_i. \tag{40}$$

Without loss, we select the outcome preferred by the provider whenever equality holds (as the provider can always slightly adjust (t,p) to break the tie to induce its preferred outcome); and select the blind-acceptance outcome if the provider is indifferent.

First-best with blind acceptance

Recall the two welfare branches

$$\mathrm{TW}^{\mathrm{BA}}(t) \;=\; B(t) - \frac{c}{1 - F(t)}, \qquad \mathrm{TW}^{\mathrm{I}}(t) \;=\; r\Big(t, \tfrac{c}{1-F(t)}\Big).$$

For TW^{I}, Proposition 1 gives quasiconcavity with unique interior maximizer t_{eff} characterized by (3). For $\mathrm{TW}^{\mathrm{BA}}$, we compute derivative

$$B'(t) \;=\; \frac{(1-\mu)\, f(t)}{1 - F(t)}\, \mathrm{MRL}(t) \;>\; 0.$$

Therefore,

$$\frac{d}{dt}\mathrm{TW}^{\mathrm{BA}}(t) \;=\; \frac{f(t)}{(1-F(t))^2}\left[(1-\mu)\int_t^{\bar{v}} (1 - F(v_i))\, dv_i - c\right].$$

Clearly, the bracket is strictly decreasing in t and the corresponding FOC coincides with the one characterizing t_{eff} in (3). Hence $\mathrm{TW}^{\mathrm{BA}}$ is strictly quasi-concave with unique interior maximizer that coincides with TW^{I}.

Since both branches peak at t_{eff}, the first-best threshold equals t_{eff} regardless of consumer strategy, and the strategy is selected by the sign of $\mathrm{TW}^{\mathrm{BA}}(t_{\mathrm{eff}}) - \mathrm{TW}^{\mathrm{I}}(t_{\mathrm{eff}})$. Since t_{eff} induces the HA outcome by Assumption A, we obtain

$$\mathrm{TW}^{\mathrm{BA}}(t_{\mathrm{eff}}) - \mathrm{TW}^{\mathrm{I}}(t_{\mathrm{eff}}) \;=\; B(t_{\mathrm{eff}}) - \tfrac{c}{1-F(t_{\mathrm{eff}})} - r_{\mathrm{eff}} \;=\; s - \mu\int_{\underline{v}}^{r_{\mathrm{eff}}} F(v_i)\, dv_i,$$

which is the cutoff stated in Lemma 4.

Usage pricing with blind acceptance

As in the proof of Proposition 2, we reformulate the problem as choosing a threshold t and a target reservation value $u \in [v_0, r_{\mathrm{eff}}]$ (implemented via $p = \Phi(u|t) - s$). Taking the consumer's branching

into account,

$$\Pi_{\text{usage}}(t,u) \;=\; \begin{cases} \Pi^{\text{BA}}_{\text{usage}}(t,u) & \text{if } B(t)+s \geq u+\Phi(u|t), \\ \Pi^{\text{I}}_{\text{usage}}(t,u) & \text{if } B(t)+s \leq u+\Phi(u|t), \end{cases} \tag{41}$$

where

$$\Pi^{\text{BA}}_{\text{usage}}(t,u) = B(t) - u - \frac{c}{1-F(t)}, \qquad \Pi^{\text{I}}_{\text{usage}}(t,u) = \frac{\Phi(u|t) - s - \frac{c}{1-F(t)}}{1-G(u|t)},$$

subject to $u \geq v_0$.

At any (t,u) on the consumer indifference set $B(t)+s = u+\Phi(u|t)$,

$$\Pi^{\text{BA}}_{\text{usage}}(t,u) - \Pi^{\text{I}}_{\text{usage}}(t,u) = \frac{-G(u|t)}{1-G(u|t)}[\Phi(u|t) - s - \frac{c}{1-F(t)}] < 0$$

given $\mu > 0$. Therefore, $\Pi_{\text{usage}}(t,u)$ is discontinuous at the branch boundary, and the provider *strictly prefers* to induce the inspection outcome. Hence for usage-pricing analysis, we select inspection outcome at the boundary.

To state the result, denote

$$(t_{\text{eff}}, u^{\circ}) \equiv \arg \max_{(t,u)\,:\,u \geq v_0} \Pi^{\text{I}}_{\text{usage}}(t,u),$$

as the *unconstrained solution* to the inspection branch, which coincides with Proposition 2. Likewise, since $\Pi^{\text{BA}}_{\text{usage}}(t,u)$ differs from $\text{TW}^{\text{BA}}(t)$ only by the constant $-u$ and is strictly decreasing in u, the analysis in the previous subsection immediately applies and we can denote

$$(t_{\text{eff}}, v_0) \equiv \arg \max_{(t,u)\,:\,u \geq v_0} \Pi^{\text{BA}}_{\text{usage}}(t,u),$$

as the *unconstrained solution* to the blind-acceptance branch. Meanwhile, denote the constrained maximum values for each branch as

$$\Pi^{\text{BA},*}_{\text{usage}} \equiv \max_{(t,u)\,:\,B(t)+s \geq u+\Phi(u|t),\; u \geq v_0} \Pi^{\text{BA}}_{\text{usage}}(t,u)$$

$$\Pi^{\text{I},*}_{\text{usage}} \equiv \max_{(t,u)\,:\,B(t)+s \leq u+\Phi(u|t),\; u \geq v_0} \Pi^{\text{I}}_{\text{usage}}(t,u).$$

Lemma 7 (Usage equilibrium with blind acceptance)**.** *Under Assumption A, the equilibrium under usage pricing satisfies the following:*

(i) if $s \leq \mu \int_{\underline{v}}^{u^{\circ}} F(v_i)\,dv_i$, then $(t_{\text{usage}}, u_{\text{usage}})$ are given by Proposition 2 and consumers inspect;

(ii) if $s > \mu \int_{\underline{v}}^{u^{\circ}} F(v_i)\,dv_i$ and $\Pi^{\text{I},*}_{\text{usage}} \geq \Pi^{\text{BA},*}_{\text{usage}}$, then $(t_{\text{usage}}, u_{\text{usage}})$ satisfies boundary constraint $B(t)+s = u+\Phi(u|t)$ and consumers inspect;

(iii) if $s > \mu \int_{\underline{v}}^{u^{\circ}} F(v_i)\,dv_i$ and $\Pi^{\text{I},*}_{\text{usage}} < \Pi^{\text{BA},*}_{\text{usage}}$, then $(t_{\text{usage}}, u_{\text{usage}}) = (t_{\text{eff}}, v_0)$ and consumers blind-accept.

Proof. Suppose

$$B(t_{\text{eff}}) + s \leq u^{\circ} + \Phi(u^{\circ}|t_{\text{eff}}), \tag{42}$$

which implies the inspection branch attains the unconstrained solution $(t_{\text{eff}}, u^{\circ})$ and induces the

HA-regime by Proposition 2. Therefore,

$$\begin{aligned}
\Pi^{\mathrm{I},*}_{\mathrm{usage}} &= \Pi^{\mathrm{I}}_{\mathrm{usage}}(t_{\mathrm{eff}}, u^{\circ}) \\
&\geq \Pi^{\mathrm{I}}_{\mathrm{usage}}(t_{\mathrm{eff}}, v_0) && \text{(definition of unconstrained solution } (t_{\mathrm{eff}}, u^{\circ})\ ) \\
&= \frac{\Phi(v_0|t_{\mathrm{eff}}) - s - \frac{c}{1-F(t_{\mathrm{eff}})}}{1-\mu F(v_0)} && \text{(high acceptance)} \\
&> B(t_{\mathrm{eff}}) - v_0 - \frac{c}{1-F(t_{\mathrm{eff}})} && \text{(by condition (42) and } \mu F(v_0) > 0) \\
&= \Pi^{\mathrm{BA}}_{\mathrm{usage}}(t_{\mathrm{eff}}, v_0) \geq \Pi^{\mathrm{BA},*}_{\mathrm{usage}}. && \text{(definition of unconstrained solution } (t_{\mathrm{eff}}, v_0)\ )
\end{aligned}$$

Moreover, given HA regime, (42) is equivalent to $s \leq \mu \int_{\underline{v}}^{u^{\circ}} F(v_i)\, dv_i$ as in the lemma statement.

If (42) fails, i.e., $s > \mu \int_{\underline{v}}^{u^{\circ}} F(v_i)\, dv_i$, then it has two implications. First, the boundary constraint on the inspection branch binds; Second, since $u^{\circ} \geq v_0$, we must have $s > \mu \int_{\underline{v}}^{v_0} F(v_i)\, dv_i$, i.e., $B(t_{\mathrm{eff}}) + s > v_0 + \Phi(v_0|t_{\mathrm{eff}})$. Therefore, the blind-acceptance branch attains its unconstrained solution (t_{eff}, v_0). Whichever branch prevails then depends on the profit comparison stated in the lemma. □

Subscription pricing with blind acceptance

Under subscription pricing $p = 0$, and the binding participation constraint gives $T(t) = \max\{B(t),\, r(t,0)\} - v_0$. Taking the consumer's branching into account,

$$\Pi_{\mathrm{subs}}(t,u) \;=\; \begin{cases} \Pi^{\mathrm{BA}}_{\mathrm{subs}}(t,u) & \text{if } B(t) \geq r(t,0), \\ \Pi^{\mathrm{I}}_{\mathrm{subs}}(t,u) & \text{if } B(t) \leq r(t,0), \end{cases}$$

where

$$\Pi^{\mathrm{BA}}_{\mathrm{subs}}(t) = B(t) - v_0 - \frac{c}{1-F(t)}, \qquad \Pi^{\mathrm{I}}_{\mathrm{subs}}(t) = r(t,0) - v_0 - \frac{c}{1-F(t)}\,\frac{1}{1-G(r(t,0)|t)}.$$

At any t on the indifference set $B(t) = r(t,0)$, the lump-sum fees coincide across configurations, so

$$\Pi^{\mathrm{BA}}_{\mathrm{subs}}(t) - \Pi^{\mathrm{I}}_{\mathrm{subs}}(t) \;=\; \frac{c}{1-F(t)}\Big(\frac{1}{1-G(r(t,0)|t)} - 1\Big) \;>\; 0.$$

The provider strictly prefers blind acceptance on the indifference set. Hence for subscription-pricing analysis, we select the BA outcome at the boundary.

Again, $\Pi^{\mathrm{BA}}_{\mathrm{subs}}$ differs from $\mathrm{TW}^{\mathrm{BA}}$ only by the constant $-v_0$, so its unconstrained maximizer is t_{eff}. To state the result, denote constrained maximum values for each branch as $\Pi^{\mathrm{BA},*}_{\mathrm{subs}} \equiv \max_{B(t)\geq r(t,0)} \Pi^{\mathrm{BA}}_{\mathrm{subs}}(t)$ and $\Pi^{\mathrm{I},*}_{\mathrm{subs}} \equiv \sup_{B(t)<r(t,0)} \Pi^{\mathrm{I}}_{\mathrm{subs}}(t)$.

Lemma 8 (Subscription equilibrium with blind acceptance)**.** *Under Assumption A, the equilibrium under subscription pricing t_{subs} satisfies the following:*

(i) if $B(t_{\mathrm{eff}}) \geq r(t_{\mathrm{eff}}, 0)$, then $t_{\mathrm{subs}} = t_{\mathrm{eff}}$ and consumers blind-accept;

(ii) if $B(t_{\mathrm{eff}}) < r(t_{\mathrm{eff}}, 0)$ and $\Pi^{\mathrm{BA},*}_{\mathrm{subs}} \geq \Pi^{\mathrm{I},*}_{\mathrm{subs}}$ then t_{subs} solves $B(t) = r(t,0)$ and consumers blind-accept;

(iii) if $B(t_{\text{eff}}) < r(t_{\text{eff}}, 0)$ *and* $\Pi^{\text{BA},*}_{\text{subs}} < \Pi^{\text{I},*}_{\text{subs}}$*, then* t_{subs} *is given by Proposition 3 and consumers inspect.*

Proof. This follows from the same reasoning as Lemma 7: if $B(t_{\text{eff}}) \geq r(t_{\text{eff}}, 0)$ then the BA branch attains its unconstrained solution and it is strictly higher than $\Pi^{\text{I},*}_{\text{subs}}$. When $B(t_{\text{eff}}) < r(t_{\text{eff}}, 0)$, then we compare the maximum of the two branches directly. □

Illustrations

Figure 8 illustrates the equilibrium threshold chosen by the provider, parallel to Figure 1 of the main text. In the first panel ($s = 0.05$, $\mu = 0.3$), only usage pricing entails blind acceptance (thick line), whereas first-best and subscription pricing still entail the inspection outcome (thin line) as in the baseline. In the second panel, the inspection cost is higher and the signal more precise (lower μ), so all three models entail blind acceptance and the equilibrium thresholds coincide at t_{eff}.[2]

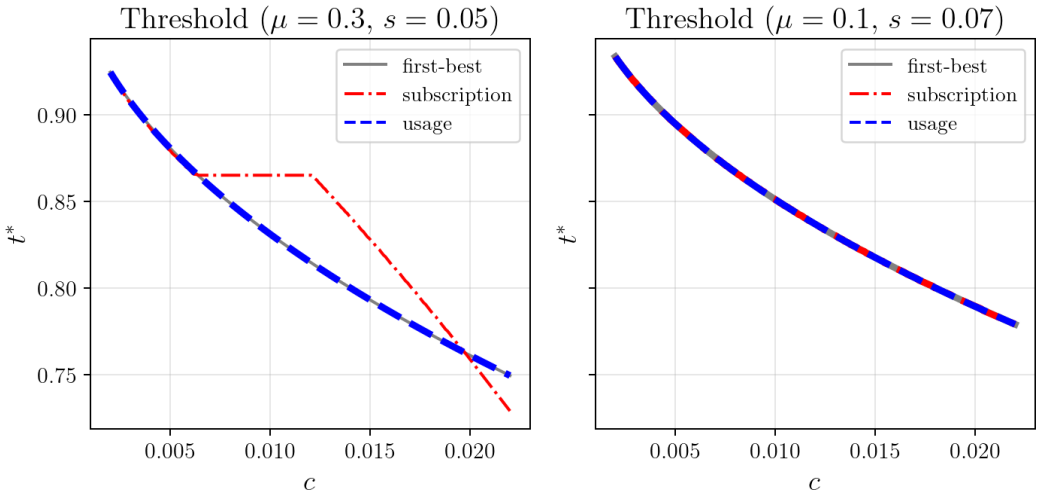

Figure 8: Equilibrium inner-search thresholds across regimes when blind acceptance is available. The parameters are $F \sim U[0, 1]$ and $s_0 = 0.1$, with (μ, s) indicated on each panel. Line thickness indicates the equilibrium consumer behavior: a thick segment marks that the corresponding regime entails blind acceptance, while a thin segment marks inspection.

C.3 Details of sponsored options

This appendix describes the omitted details for the analysis in Section 6.3. In what follows, we let α^R denote the expected number of queries, n^R_O and n^R_S the expected numbers of organic and sponsored displays (by definition, $n^R_O + n^R_S = \alpha^R$), and β^R_S the total probability that a sponsored recommendation is displayed and accepted.

[2]The corresponding profit plots are omitted, as the pattern is qualitatively similar to that of Figure 1 of the main text, except that whenever all three models entail blind acceptance the three profit curves coincide.

Opacity

The consumer does not observe whether the displayed recommendation is organic or sponsored, and perceives each recommendation as drawn from the mixture

$$H(v \,|\, t_O, t_S) \;=\; (1-\lambda)\, G_O(v|t_O) + \lambda\, G_S(v|t_S),$$

and the analysis collapses to the single-pool blind-acceptance extension (Section 6.2) applied to H. Define the mixed blind-accept value

$$B(t_O, t_S) \;\equiv\; \int_{\underline{v}}^{\bar{v}} v_i \, dH(v_i|t_O, t_S)$$

Recall that in equilibrium the consumer's querying strategy is exactly one of the following:

- **(BA) Blind-accept.** Stop at first query. Per consumer: $\alpha = 1$, $n_S = \lambda$, $n_O = 1-\lambda$, $\beta_S = \lambda$. Consumer utility is $U = B(t_O, t_S) - p$.
- **(I) Always inspect.** Reservation r solves

$$\int_r^{\bar{v}} \big(1 - H(v_i|t_O, t_S)\big)\, dv_i = s + p.$$

  Per query, acceptance probability is $1 - H(r|t_O, t_S)$, so $\alpha = 1/[1-H(r)]$, $n_S = \lambda\alpha$, $n_O = (1-\lambda)\alpha$, $\beta_S = \lambda\big[1 - G_S(r|t_S)\big]/[1-H(r)]$. Consumer utility is $U = r$.

Then, for any given t_O, t_S and pricing chosen by the provider, the BA regime prevails if $B(t_O, t_S) - p \geq r$, and the I regime prevails otherwise.

The provider's optimization problem is

$$\Pi_{\text{opacity}} = \; T \;+\; p\,\alpha \;+\; A\, n_S \;+\; a\,\beta_S \;-\; \kappa_O\, n_O \;-\; \kappa_S\, n_S$$

where $T = U - v_0$ and $p = 0$ under subscription-pricing; and $T = 0$ and p is an optimization variable under usage-pricing. As noted above, α, n_S, n_O, and β_S can take different forms depending on which regime prevails for the given t_O, t_S and pricing chosen by the provider.

Transparency

The consumer observes the type before deciding whether to blind-accept, inspect, or query again. Let

$$B_j(t_j) \equiv \int_{\underline{v}}^{\bar{v}} v \, dG_j(v|t_j)$$

denote the blind-accept value of a type-j ($j \in \{O, S\}$) recommendation. Let

$$I_j(r; t_j) \equiv \int_{\underline{v}}^{\bar{v}} \max\{v, r\}\, dG_j(v|t_j) = r + \int_r^{\bar{v}} [1 - G_j(v|t_j)]\, dv$$

and

$$q_j(r; t_j) \equiv G_j(r|t_j) = \frac{F(r) - F(t_j)}{1 - F(t_j)}\, \mathbf{1}_{\{r \geq t_j\}}$$

denote, respectively, the value of inspecting a type-j recommendation and the probability of rejection after inspection, when the continuation value is r.

The consumer's stationary value of querying satisfies

$$r = -p + (1-\lambda)M_O(r;t_O) + \lambda M_S(r;t_S), \tag{43}$$

where

$$M_O(r;t_O) \;=\; \max\{B_O(t_O),\;\; r,\;\; I_O(r;t_O) - s\}, \tag{44}$$

$$M_S(r;t_S) \;=\; \max\{B_S(t_S),\;\; r,\;\; I_S(r;t_S) - s\}. \tag{45}$$

The three terms inside each maximum correspond to blind acceptance (BA), re-querying without inspection (Q), and inspection (I). Comparing the slopes of both sides shows that (43) admits a unique fixed point r: the left-hand side has slope 1 in r, while the right-hand side has slope strictly less than 1 (outside the Q-Q branch, which is ruled out below).

Let $\sigma_O \in \{\text{BA}, \text{Q}, \text{I}\}$ denote the maximizing branch in (44) and $\sigma_S \in \{\text{BA}, \text{Q}, \text{I}\}$ the maximizing branch in (45). Label regimes by σ_O-σ_S (consumer decision upon receiving an organic and a sponsored recommendation respectively). Q-Q is ruled out: it would require the consumer to never stop, which is inconsistent with finite r in (43). Then, the solution to (43) determines which regime arises.

We now describe the key endogenous objects that arise in each regime. Conditional on a recommendation being organic, the probability that a consumer stops querying is

$$\rho_O(\sigma_O, r; t_O) = \begin{cases} 1 & \text{if } \sigma_O = \text{BA} \\ 0 & \text{if } \sigma_O = \text{Q} \\ 1 - q_O(r;t_O) & \text{if } \sigma_O = \text{I} \end{cases},$$

and analogously $\rho_S(\sigma_S, r; t_S)$ for sponsored. The per-query stopping probability is

$$\rho^\star \;=\; (1-\lambda)\,\rho_O \;+\; \lambda\,\rho_S.$$

Each query independently shows organic with probability $1-\lambda$ and sponsored with probability λ. Therefore, standard geometric stopping delivers:

$$\begin{aligned} \alpha &= \frac{1}{\rho^\star} && \text{(expected total queries)} \\ n_O &= \frac{1-\lambda}{\rho^\star}, \qquad n_S = \frac{\lambda}{\rho^\star} && \text{(expected organic / sponsored displays)} \\ \beta_O &= \frac{(1-\lambda)\,\rho_O}{\rho^\star}, \qquad \beta_S = \frac{\lambda\,\rho_S}{\rho^\star} && \text{(probability of accepting organic / sponsored).} \end{aligned}$$

Note $\beta_O + \beta_S = 1$ and $n_O + n_S = \alpha$ in any regime.

The provider's profit takes the same form as in the opacity case,

$$\Pi_{\text{transparency}} = T + p\,\alpha + A\,n_S + a\,\beta_S - \kappa_O\,n_O - \kappa_S\,n_S,$$

with $\alpha, n_S, n_O, \beta_S$ now given by the transparency renewal expressions above.

The numerical implementation solves the fixed points (43), classifies the induced stopping regime, and solves the profit maximization. The results are reported in Section 6.3.

C.4 Impression advertising

We now consider a variant of the baseline that incorporates impression ads, i.e., ad revenue that scales with the total number of user queries. In the baseline model, conversion-based ad revenues have no impact on the equilibrium outcome. This is because consumers eventually accept an option in equilibrium, and so such revenues simply enter the profit function as a constant additive component. To capture this with a minimal departure from our baseline model, we modify the provider's per-query cost to:

$$\frac{c}{1-F(t)} \; - \; A,$$

where $A \geq 0$ is the advertising revenue collected each time a consumer queries. When $A = 0$ we recover the baseline model. We treat A as the economic value generated from advertising, and so A enters total welfare as a genuine social benefit in our efficiency benchmark.

Efficiency benchmark under impression ads. The efficient per-query fee is $p = \frac{c}{1-F(t)} - A$, internalizing the net cost. Following the same analysis as in Proposition 1:

(i) If $\mu c < (1-\mu)(s-A)$: the efficient threshold t_{eff} is the unique solution to (3), and it induces a high-acceptance regime.

(ii) If $\mu c \geq (1-\mu)(s-A)$: the efficient threshold $t_{\text{eff}} = \underline{v}$, and it induces a low-acceptance regime.

Compared to the baseline setting (Proposition 1), advertising revenue shrinks the parameter range where the efficient threshold entails a high-acceptance regime, i.e., $\mu c < (1-\mu)(s-A)$ becomes harder to hold as A rises. This is intuitive because advertising revenue makes frequent querying (low-acceptance regime) relatively more efficient. Nonetheless, within each regime, the efficient threshold is identical to the baseline.

AI provider choices. Parallel to Assumption A, throughout this subsection we assume:[3]

$$\mu c < (1-\mu)(s-A) \quad \text{and} \quad r(t_{\text{eff}}, \frac{c}{1-F(t_{\text{eff}})} - A) > v_0. \tag{46}$$

The assumption implies t_{eff} is independent of A. We now examine how raising A affects the choices of the AI provider.

Corollary 3 (Impression ads and thresholds)**.** *Suppose (46) holds. Under usage pricing, $t_{\text{usage}} = t_{\text{eff}}$ is independent of A. Under subscription pricing, there exists a cutoff $\hat{t}_{\text{subs}} \in (\underline{v}, \bar{v})$ such that:*

1. *if $t_{\text{subs}} > \hat{t}_{\text{subs}}$ at some $A = A' \geq 0$, then t_{subs} is increasing in A for all $A \geq A'$;*
2. *if $t_{\text{subs}} < \hat{t}_{\text{subs}}$ at some $A = A' \geq 0$, then t_{subs} is decreasing in A for all $A \geq A'$.*

Proof. **(Corollary 3).** We analyze first-best, usage pricing, and subscription pricing in turn.

Total welfare. Given $\text{TW}(t) = r\big(t, \frac{c}{1-F(t)} - A\big)$, define the boundary cutoff as

$$r(\hat{t}_{\text{eff}}, \frac{c}{1-F(\hat{t}_{\text{eff}})} - A) = \hat{t}_{\text{eff}}.$$

[3] Assumption A neither implies nor is implied by (46): a higher A tightens the first part but relaxes the second.

In the HA branch ($t \geq \hat{t}_{\text{eff}}$), derivatives from Lemma 2 give:

$$\frac{d}{dt}\text{TW}(t) = \frac{f(t)}{[1-\mu F(r)][1-F(t)]^2}\Big[(1-\mu)\int_t^{\bar{v}}[1-F(x)]\,dx - c\Big].$$

The solution to the FOC, denoted as t°, is identical to (3) in the baseline. By the same feasibility argument as in Proposition 1, $t^\circ \geq \hat{t}_{\text{eff}}$ only if $(1-\mu)(s-A) \geq \mu c$. In the LA branch ($t < \hat{t}_{\text{eff}}$),

$$\frac{d}{dt}\text{TW}(t) = \frac{f(t)}{(1-F(r))[1-\mu F(t)]^2}\big[(1-\mu)(s-A) - \mu c\big],$$

which has constant sign. Combining the two branches as in the proof of Proposition 1, $t_{\text{eff}} = t^\circ$ when $(1-\mu)(s-A) > \mu c$, and $t_{\text{eff}} = \underline{v}$ when $(1-\mu)(s-A) < \mu c$. Without loss of generality, we select $t_{\text{eff}} = \underline{v}$ at the boundary.

Usage pricing. We denote $\tilde{r}_{\text{eff}} \equiv r(t_{\text{eff}}, \frac{c}{1-F(t_{\text{eff}})} - A)$. Similar to the proof of Proposition 2, we fix target reservation value using $u = r(t,p) \in [v_0, \tilde{r}_{\text{eff}}]$, which is non-empty by assumption (46). Then, we optimize with respect to t:

$$\tilde{\Pi}_{\text{usage}}(t,u) = \Pi_{\text{usage}}(t,u) + \frac{A}{1-G(u|t)}.$$

In the HA branch ($t \geq u$), $1-G(u|t) = 1-\mu F(u)$ is constant in t, and so $t_{\text{usage}} = t^\circ$ by Proposition 2. Moreover, by assumption (46), $t^\circ = t_{\text{eff}} > \tilde{r}_{\text{eff}} \geq u$, so the solution does not violate the boundary constraint. In the LA branch ($t < u$), a direct calculation gives

$$\begin{aligned}
\tilde{\Pi}_{\text{usage}}(t) &= \frac{sF(t)(1-\mu) - c + A(1-F(t))}{(1-F(u))(1-\mu F(t))}. \\
\frac{d}{dt}\tilde{\Pi}_{\text{usage}} &= \frac{f(t)}{(1-F(u))(1-\mu F(t))^2}\Big[(1-\mu)(s-A) - \mu c\Big] > 0
\end{aligned}$$

by assumption (46). Finally, since $\tilde{r}_{\text{eff}}$ is the highest achievable total welfare, non-negativity of profit requires $u \leq \tilde{r}_{\text{eff}}$, as claimed; similarly, the participation constraint implies $u \geq v_0$.

Subscription pricing. We know $T = r(t,0) - v_0$ continues to hold. Then,

$$\tilde{\Pi}_{\text{subs}}(t) = \Pi_{\text{subs}}(t) + A \cdot \alpha(t,0).$$

The cross derivative is $\frac{\partial^2 \tilde{\Pi}_{\text{subs}}}{\partial t\,\partial A} = \frac{\partial}{\partial t}\alpha(t,0)$. Fix some $A = A'$. By the implicit function theorem, whenever t_{subs} is interior,

$$\frac{\partial}{\partial t}\alpha(t_{\text{subs}},0)|_{A=A'} > 0 \Rightarrow \frac{d}{dA}t_{\text{subs}}|_{A=A'} > 0.$$

Recall cutoff $\hat{t}_{\text{subs}}$, as defined in (17), is such that $t > \hat{t}_{\text{subs}} \Rightarrow t > r(t,0)$ (HA regime), and vice versa. So, Lemma 2 implies

$$t_{\text{subs}}|_{A=A'} > \hat{t}_{\text{subs}} \Rightarrow \frac{\partial}{\partial t}\alpha(t_{\text{subs}},0)|_{A=A'} > 0.$$

Any further increase in A raises t_{subs} while leaving $\hat{t}_{\text{subs}}$ unchanged, so $t_{\text{subs}} > \hat{t}_{\text{subs}}$ continues to hold for all $A \geq A'$. The opposite case of $t_{\text{subs}}|_{A=A'} < \hat{t}_{\text{subs}}$ is established by the same argument and

is therefore omitted. Note this proof is silent when $t_{\text{subs}}|_{A=A'} = \hat{t}_{\text{subs}}$ as the two-sided derivative is not well-defined at the kink point. □

Intuitively, Corollary 3 reflects that the impression ad revenue incentivizes the provider to ensure a high volume of queries. Under usage pricing, the equilibrium is always a high-acceptance regime, so the provider can only increase query volume by raising the threshold t (see Lemma 2), which continues to align its incentive with efficiency. Under subscription pricing, the cutoff $\hat{t}_{\text{subs}}$ is defined such that $t_{\text{subs}} > \hat{t}_{\text{subs}}$ implies a high-acceptance regime, whereas $t_{\text{subs}} < \hat{t}_{\text{subs}}$ implies a low-acceptance regime. In the former, the provider again raises t to generate more queries; in the latter, it does so by lowering t.

Meanwhile, Corollary 4 below shows that a subscription-pricing provider benefits more from advertising revenue than a usage-pricing provider. This is intuitive because advertising revenue scales with the number of user queries, which is naturally higher under subscription pricing.

Corollary 4 (Impression ads and profit)**.** *The equilibrium profits* Π^*_{subs} *and* Π^*_{usage} *are strictly increasing in* A*. Moreover, if* μ *is sufficiently small, then*

$$\left[\frac{d}{dA}\Pi^*_{\text{subs}} - \frac{d}{dA}\Pi^*_{\text{usage}}\right]_{A=0} > 0.$$

That is, starting from $A = 0$*, a marginal increase in advertising revenue* A *expands the parameter region where* $\Pi^*_{\text{subs}} > \Pi^*_{\text{usage}}$*.*

Proof. **(Corollary 4).** For each model $\omega \in \{\text{usage}, \text{subs}\}$, the envelope theorem implies $\frac{d}{dA}\Pi^*_\omega$ equals the equilibrium number of queries $\alpha_\omega > 0$. It remains to compare α_{usage} and α_{subs} at $A = 0$. Under usage pricing, when $\mu \to 0$, (16) implies $u_{\text{usage}} \to v_0$, and so $\alpha_{\text{usage}} = \frac{1}{1-\mu F(v_0)}$. Under subscription pricing, by the definition of the cutoff $\hat{t}_{\text{subs}} \in (\underline{v}, \bar{v})$ from (17) and Lemma 2,

$$\frac{1}{1 - G(r(t_{\text{subs}}, 0)|t_{\text{subs}})} = \frac{1}{1 - \mu F(r(t_{\text{subs}}, 0))} > \frac{1}{1 - \mu F(r(\hat{t}_{\text{subs}}, 0))} \qquad \text{if } t_{\text{subs}} > \hat{t}_{\text{subs}};$$

$$\frac{1}{1 - G(r(t_{\text{subs}}, 0)|t_{\text{subs}})} > \frac{1}{1 - G(r(\hat{t}_{\text{subs}}, 0)|\hat{t}_{\text{subs}})} = \frac{1}{1 - \mu F(r(\hat{t}_{\text{subs}}, 0))} \qquad \text{if } t_{\text{subs}} < \hat{t}_{\text{subs}}.$$

Therefore, it remains to prove $r(\hat{t}_{\text{subs}}, 0) > v_0$. To do so, we note by definition $r(\hat{t}_{\text{subs}}, 0) = \hat{t}_{\text{subs}}$, where

$$\Phi(\hat{t}_{\text{subs}}|\hat{t}_{\text{subs}}) = s \leq s_0 = \Phi(v_0|\underline{v}).$$

By the properties of $\Phi(\cdot|\cdot)$ shown in the proof of Lemma 1, we conclude $\Phi(\hat{t}_{\text{subs}}|\hat{t}_{\text{subs}}) < \Phi(v_0|\hat{t}_{\text{subs}})$, so $\hat{t}_{\text{subs}} > v_0$. □