Implications of LLM Integration in Advertising *

Eduardo M. Azevedo† Alessandro Sellitto‡ Pinar Yildirim§

This version: October 23, 2026

Abstract

We examine how richer query information changes the revenue economics of search advertising. We model the move from keyword search to conversational search as a refinement of the platform's information about query-level click-through rates and ask how this refinement affects expected revenue under different selling rules. We find that, under a fixed weighted second-price auction, finer information weakly improves matching but has an ambiguous effect on revenue. This is because more precise queries can thin out competition within realized auctions. When the platform can adapt the mechanism to the information it observes, finer information weakly raises expected revenue. With constant demand elasticities, the revenue-maximizing auction is a weighted second-price auction with tuned weights and reserves, an auction format major platforms use. Conversational search also raises questions regarding the optimal timing of ads in a conversation. We argue that a platform may prefer to wait before showing an ad, because later turns can reveal more about what the user wants, although the user may leave before then.

Keywords: sponsored search; auctions; large language models; online advertising; mechanism design; information.

JEL Codes: D44, D82, M37.

*We are grateful to Juan Andres Russy for excellent research assistance. We also thank Francesco Decarolis and seminar participants at Institut Mines-Télécom Business School for valuable feedback.

†The Wharton School, University of Pennsylvania, eazevedo@wharton.upenn.edu.

‡Institut Mines-Télécom Business School, University of Milan, and University of Pavia, alessandro.sellitto@imt-bs.eu.

§The Wharton School, Marketing Department & Department of Economics, University of Pennsylvania and NBER, pyild@wharton.upenn.edu.

1 Introduction

Whether large language models (LLMs) should be monetized through advertising has become a prominent debate with growing commercial significance. In early 2026, OpenAI reported more than 900 million weekly active users for ChatGPT, yet only about six percent were estimated to hold a paid subscription (Malik, 2026; Kafka, 2026). The company subsequently began testing personalized advertising and projected advertising revenues of $2.5 billion for the year 2026, rising to approximately $100 billion by 2030 (OpenAI, 2026d; Fried, 2026; Reuters, 2026). ChatGPT Ads had reached a reported annualized revenue run rate of $1 billion by August 2026, indicating a rapid development of this new monetization channel (OpenAI, 2026a). Other firms adopted different approaches. Anthropic has publicly committed to keeping Claude ad-free, arguing that advertising in conversational AI could undermine user trust. The company reiterated its position through a Super Bowl campaign with the tagline "Ads are coming to AI. But not to Claude" (Anthropic, 2026a; Griffiths, 2026; Preston, 2026). Alphabet has developed advertising formats for AI-powered search, integrating ads into search output that incorporates AI-generated answers (Google, 2026a; Nolan, 2026).

One prominent concern is that conversational, answer-first interfaces may disrupt established search advertising models. Advertising executives have questioned what these interfaces mean for a model in which users navigate a series of links to assess which products best fit their needs. Frederick Vallaeys, a former Google advertising executive, stated that "for the first time in 20 years, I actually believe that keywords are dead" (Barnett and Hiebert, 2025). Google users who saw an AI summary clicked on a traditional search result in 8% of visits, compared with 15% when no summary appeared (Chapekis and Lieb, 2025). Financial press coverage has likewise emphasized the "sharp fall in web traffic" associated with AI summaries and the potential threat these interfaces pose to the advertising model of the open web (Thomas, 2025; Bradshaw et al., 2025).

In this paper we focus on a single yet consequential margin along which conversational search environments differ from traditional keyword-search environments: consumers provide substantially richer information in their queries. Consider, for instance, a consumer searching for car insurance. Under the keyword regime, the consumer communicates intent through a short textual cue such as "car insurance," and is shown ads from advertisers who bid on those keywords in a sponsored-search auction (Edelman et al., 2007; Varian, 2007). Conversational search interfaces, in contrast, elicit more granular information regarding the intent of the consumer. Queries from the public LMSYS-Chat corpus include insurance-related requests about finding "an agreed value policy" for "my '85 Ford Mustang that I use at high performance driving events, but also drive on the street," or asking under what conditions a totaled vehicle covered by "GAP insurance" would have its loan paid off.[1] These detailed queries, together with consumers' earlier queries, allow for a finer partition of consumer segments and more precise consumer targeting than advertisers can obtain from keyword queries alone.

[1]LMSYS-Chat-1M is a public, license-gated dataset released by the Large Model Systems Organization. The corpus contains one million user conversations with 25 large language models, collected from the Vicuna demo and Chatbot Arena between April and August 2023. The dataset was gathered from approximately 210, 000 unique IP addresses and includes conversation text, the model identifier, detected language, and moderation labels. The dataset is publicly available through the Hugging Face repository (Zheng et al., 2024; LMSYS Organization, 2023).

We develop a model of a platform that can display at most one advertisement for each query, with advertisers' private values per click converted into query-specific values per impression through click-through rates, and derive results under two informational regimes. In the *low-information regime*, closely resembling the keyword era, consumers present the platform with a short textual query with limited information about the object of search. In the *high-information regime*, corresponding to LLM-based or conversational search, the platform observes finer partitions of the low-information query and conducts a separate auction for each. We consider two auction regimes. In the fixed auction regime, the platform is limited to a weighted second-price auction, with weights equal to click-through rates and a common reserve, as in Edelman et al. (2007) and Varian (2007). In the flexible auction regime, the platform can select an arbitrary mechanism for each query, as in Myerson (1981). The ability to adjust auction parameters allows the platform to use richer query information in determining ad allocation and prices.

We discuss three key findings. First, we show that the revenue effect of richer query information depends on the auction regime. With a fixed auction, more information can reduce competition, because a refined query can have one obvious best advertiser, so revenue may rise or fall (Levin and Milgrom, 2010). In a flexible auction, when the platform can adapt the mechanism to each query, more precise information weakly raises expected revenue. The intuition follows the characterization of the revenue-maximizing auction in Myerson (1981), and the result is closely related to Fu et al. (2012). Our contribution is to apply it to the move from keyword to conversational search. More detailed query information enters through query-specific click-through rates, which determine how much an impression is worth to each advertiser. A richer query can therefore change both the advertiser who is the best match to a query and the price the platform can charge for an impression.

Second, we show that when advertisers' demand elasticities are constant, the optimal auction can be implemented as a weighted second-price auction with data-tuned weights and reserves, a format that most sponsored-search platforms use. Sponsored-search platforms use quality scores, predicted click-through rates, bid adjustments, and reserve prices to determine which ads are shown and how much advertisers pay. Google provides an example of how the parameters can be adjusted in practice. Although Google uses a generalized second-price auction, its pricing rules are not fixed as in the early sponsored-search models. Documents disclosed in *U.S. and Plaintiff States v. Google LLC* describe several pricing adjustments, including format pricing, squashing, and reserves (U.S. Department of Justice, 2023, trial exhibit UPX0509; see Section 5), through which the platform can change how advertisers are ranked and how much they pay for different queries.

Third, we extend the analysis to the timing of ads within a conversation. If the platform can display only one ad, it may prefer to wait for additional messages rather than advertise immediately. Later messages may provide more information about the user's needs and increase the revenue from the auction. However, waiting also carries the risk that the user leaves before an ad is shown. Under a flexible auction, more detailed information can make waiting profitable, particularly when the user is likely to continue the conversation. Under a fixed weighted second-price auction, information itself can favor showing the ad early. A more specific turn may leave one advertiser as the clear best match, which reduces competition and lowers the price, so the platform may advertise before the user's intent is clear even when no user would leave.

Our paper contributes first to the industry debate over advertising in conversational search. Much of this debate centers on the possibility that conversational, answer-first interfaces reduce clicks and thereby may harm search advertising. We argue a countervailing force may mitigate these concerns. While conversational interfaces can reduce the "list-of-links" browsing margin, they also give platforms richer information about the commercial intent and preferences of consumers. When refined information can be fed into a flexible auction, search platforms can sustain and may even grow advertising revenues.

Second, we contribute to the literature on information in auctions. Our study relates to Eső and Szentes (2007), who show that when the seller can choose both what information to disclose and how to sell the object, full disclosure can be revenue-maximizing and implementable through a handicap auction. Existing work also shows that finer information can improve allocation while leaving seller revenue ambiguous under fixed auction formats (Board, 2009; Ganuza and Penalva, 2010; Hummel and McAfee, 2016; Levin and Milgrom, 2010). Our contribution is to show how this distinction operates in sponsored search when richer queries refine click-through-rate estimates. Under a fixed weighted second-price auction, refining the query can thin competition and may raise or lower revenue. Larsen and Proserpio (2026) document a pattern consistent with this mechanism around Google's BERT rollout. Better query understanding lowers cost per click on long, context-rich queries and raises it on short ones. The cost decline on long queries is consistent with consumer segmentation thinning ad competition, so that fewer advertisers end up being a close match for each query. Under a flexible query-level mechanism, expected revenue is the expected virtual surplus from the Myerson auction; because virtual surplus is convex in the click-through-rate vector, the refinement weakly raises revenue (Myerson, 1981; Fu et al., 2012; Arefeva and Meng, 2021). So, the value of conversational query data depends on the auction regime. Richer prompts are most valuable when a platform can adjust its reserves, weights, and other auction parameters at the query level.

Third, we contribute to the literature on sponsored-search and position auctions (Edelman et al., 2007; Varian, 2007; Lahaie and Pennock, 2007; Athey and Ellison, 2011). A growing literature studies how auction design changes when advertising is mediated by LLMs and generated content (Feizi et al., 2023; Dubey et al., 2024; Hajiaghayi et al., 2024; Soumalias et al., 2024; Balseiro et al., 2025; Zhao et al., 2025; Bhawalkar et al., 2025). These papers show that conversational and generative interfaces create retrieved ads, sponsored questions, and advertiser influence over generated text. We focus on conversational search improving platform information about match values before the advertising opportunity is sold. Related work examines the timing of advertising within conversations. Banchio et al. (2024) study how first-price and second-price auctions generate different incentives for when to run an auction. More closely related to our dynamic setting, Alaei et al. (2026) study optimal advertising timing when the platform gradually learns about user preferences while facing the risk of user departure, and develop approximately optimal incentive-compatible mechanisms. Our analysis is complementary to both papers. Rather than focusing on the design of a dynamic mechanism, we examine how the value of waiting depends on the auction regime. In particular, as later turns reveal more precise information about query-level click-through rates, expected auction revenue can increase or decrease under a fixed weighted second-price auction, whereas it weakly increases under a flexible mechanism.

Finally, we contribute to research on online advertising. Studies have examined how targeting affects advertising effectiveness (Goldfarb and Tucker, 2011) and advertising prices across media (Bergemann

and Bonatti, 2011). Others examine concentration and buyer power in online advertising (Decarolis and Rovigatti, 2021) and bidding strategies in repeated auctions (Balseiro and Gur, 2019). We study how conversational search provides platforms with more detailed information about users and how this information affects ad allocation and auction revenue. Our results also have implications for privacy. If platforms can increase revenue by adjusting their auctions to richer query information, they have stronger incentives to collect, retain, and use detailed information about users.

The rest of the paper proceeds as follows. Section 2 sets out the model, and Section 3 presents the main results. Section 4 extends the analysis to the timing of advertisements within conversations. Section 5 discusses implications, and Section 6 concludes.

2 Model

We develop a simple model to deliver our key discussion points. There is a platform, such as a search engine or conversational assistant, a finite set of advertisers $i = 1, \ldots, I$, and a finite set of granular query types $j = 1, \ldots, J$. Query type j occurs with probability m_j, with $\sum_{j=1}^{J} m_j = 1$.

Advertiser i has a private value per click u_i, drawn from distribution F_i with density f_i. We assume values are independently drawn across advertisers, and that the distributions are common knowledge. We interpret u_i as the expected return generated by a click on advertiser i's ad. For example, a car insurer that converts ad clicks into expensive, long-term policies may value a click more than one selling minimum-coverage policies.

The click-through rate (CTR) of advertiser i on query j is denoted by $\alpha_{ij} \in [0, 1]$ and is commonly known.[2] The value per impression of showing advertiser i's ad on query j is

$$v_{ij} := \alpha_{ij} u_i.$$

The platform can display at most one advertisement in response to a query and incurs a display cost $c > 0$ whenever it displays an ad. We vary the environment along two dimensions: the information available to the platform about the query and the flexibility of the auction format.

Definition 1 (Information Regimes)**.**

1. In the **Low-Information Regime**, the platform observes only an aggregate query $j = 0$ (such as "car insurance"). The platform runs a single auction.

2. In the **High-Information Regime**, the platform observes a granular query $j \in \{1, \ldots, J\}$ (such as "cheap car insurance in New York," "comprehensive car insurance," ...). The platform runs one auction per query.

Query j occurs with probability m_j and aggregate CTRs are equal: $\alpha_{i0} = \sum_{j=1}^{J} m_j \alpha_{ij}$.

Definition 2 (Auction Flexibility Regimes)**.**

[2] We treat the CTRs as common knowledge to keep the focus of the analysis on the effects of the information held by the platform.

1. In the **fixed auction regime**, the platform runs a weighted second-price auction[3] for all queries, with weights equal to the click-through rates and the same reserve price (Edelman et al., 2007).

2. In the **flexible auction regime**, the platform can select a mechanism for each query (Myerson, 1981).

In the fixed auction regime, bidders use the dominant-strategy equilibrium bid $b_i = u_i$. In the flexible auction regime, the mechanism must be Bayesian incentive compatible and individually rational.

Moving from the low-information regime to the high-information regime does not change advertisers' values, the distribution of users, or the set of ads that can be shown. It allows the platform to condition the auction on a more precise description of the query. Auction flexibility describes what the platform can do with this additional information. Under fixed auctions, the platform uses the same auction format after information improves. Under flexible auctions, the platform can redesign the mechanism for each query type. The distinction is important because better matching information (i) can improve allocation by sending impressions to advertisers with higher match values and (ii) can change the intensity of competition within each auction. The analysis we provide disentangles these forces.

3 Analysis

In this section, we provide an analysis and a comparison of the two auction regimes. We define revenue as the platform's expected revenue net of display costs, which the platform maximizes. We define welfare as the value of allocated ads minus display costs.

We first consider the fixed auction regime, in which the platform uses a weighted second-price auction whose weights equal the click-through rates and whose reserve does not change with the information available. We show that more detailed query information weakly increases welfare, while its effect on revenue can go in either direction. Better information allows the platform to match ads more accurately to users and can reduce competition when one advertiser becomes the clear winner.

We then consider the flexible auction regime, in which the platform can choose the revenue-maximizing mechanism for each query. In this case, more detailed information weakly increases expected revenue, consistent with the findings of Fu et al. (2012). We also show that, under constant elasticity, the optimal mechanism can be implemented as a weighted second-price auction with appropriately chosen weights and reserves (Proposition 4). Section 4 extends the analysis to the platform's decision of when to show an ad during a conversation.

3.1 Fixed Auction Regime

The first standard result is that finer information always increases efficiency if the reserve price is c. The intuition is that, with better information, ads can be targeted more efficiently to relevant queries.

[3] See Definition 3 in Appendix A.

Proposition 1 (With Fixed Auctions, More Information Weakly Increases Efficiency)**.** *In the fixed auction regime with reserve price c, efficiency is weakly higher in the High-Information Regime than in the Low-Information Regime.*

Proof. Efficiency from query j is

$$\mathbb{E}\left[\max\{0, u_1\alpha_{1j} - c, \ldots, u_I\alpha_{Ij} - c\}\right].$$

This is the maximum of affine functions, and thus convex in the vector $(\alpha_{ij})_{i=1}^{I}$. Jensen's inequality implies that average efficiency across the granular queries $j = 1, \ldots, J$ is weakly higher than efficiency from the aggregate query $j = 0$. □

Proposition 1 isolates the allocation effect of better query information. It indicates that finer query information improves the matching of ads to users. In the low-information regime, with an aggregate query such as "car insurance," the platform evaluates each advertiser by its average click-through rate across all users who type that query. In the high-information regime, if the platform can tell users who want cheap coverage from users who want comprehensive coverage, it can show each group the advertiser they are most likely to click on. Finer information therefore never makes the allocation worse, because the platform could always ignore it and allocate as in the low-information regime.

We next ask whether richer information also raises the platform's revenue. On the one hand, better information increases the ability to target ads. On the other hand, it thins out competition since better predictions of click-through rates may lead to auctions with an obvious winner, and thus reduce the second-highest bid.[4]

Proposition 2 (With Fixed Auctions, More Information Can Increase or Decrease Revenue)**.** *In the fixed auction regime,*

1. *If $I = 2$ and there is no reserve price, expected revenue is weakly lower in the High-Information Regime.*

2. *If $I > 2$, the effect of the information regime on expected revenue is ambiguous.*

Proof. For $I = 2$ and no reserve price, revenue is

$$\mathbb{E}\left[\min(u_1\alpha_{1j}, u_2\alpha_{2j})\right] - c.$$

The function $\min(u_1\alpha_{1j}, u_2\alpha_{2j})$ is concave in $(\alpha_{1j}, \alpha_{2j})$, since it is the pointwise minimum of linear functions. Jensen's inequality therefore implies that segmentation weakly lowers expected revenue.

For $I > 2$, the effect is ambiguous. This is proven by example in Appendix A. □

[4]The result that revenue always decreases with two bidders is essentially a special case of Board (2009), Proposition 1. Board (2009) does not consider a weighted auction, but the same logic applies.

3.2 Flexible Auction Regime

In the flexible auction regime, finer information weakly increases platform revenue. Throughout this subsection, we assume that each F_i is regular: advertiser i's virtual value per click,

$$\tilde{\psi}_i(u_i) := u_i - \frac{1 - F_i(u_i)}{f_i(u_i)},$$

is strictly increasing in u_i. To show the result, we first characterize the optimal auction.

Lemma 1 (Optimal Auction; Myerson (1981))**.** *Suppose that each F_i is regular. Then the following dominant-strategy incentive-compatible auction maximizes expected revenue from query j:*

1. *Bidders announce their value per click u_i.*
2. *The platform computes the effective bids*

$$\alpha_{ij}\left(1 - \frac{1}{\epsilon_i(u_i)}\right) u_i,$$

   where the price-per-click elasticity of demand is

$$\epsilon_i(u_i) = \frac{u_i f_i(u_i)}{1 - F_i(u_i)}.$$

3. *The winner is the bidder with the highest effective bid, provided that this bid exceeds the display cost c. Otherwise, the impression is not allocated.*
4. *The winner pays the minimum per-click bid necessary to win. Let T_{-i} be the threshold effective bid, defined as the maximum of the display cost c and the second-highest effective bid. The winning bidder i pays a price per click u_i^* that solves:*

$$\alpha_{ij}\left(1 - \frac{1}{\epsilon_i(u_i^*)}\right) u_i^* = T_{-i}.$$

The proof is in Appendix A. The lemma is the Myerson (1981) optimal auction applied to value per impression rather than value per click. The only difference from the standard single-object auction is that advertiser i's value for the impression is scaled by the query-specific click-through rate α_{ij}. The platform therefore ranks advertisers by virtual value per impression, not by value per click.

Proposition 3 (More Information Increases Revenue)**.** *In the flexible auction regime, expected revenue is weakly higher in the High-Information Regime than in the Low-Information Regime.*

Proof. Myerson (1981) shows that expected revenue from query j is

$$\mathbb{E}\left[\max\left\{0, \alpha_{1j}\tilde{\psi}_1(u_1) - c, \ldots, \alpha_{Ij}\tilde{\psi}_I(u_I) - c\right\}\right],$$

where $\tilde{\psi}_i$ is the virtual value per click defined above. This is the maximum of affine functions, and thus convex in the vector $(\alpha_{ij})_{i=1}^{I}$. Jensen's inequality implies that average revenue across the granular queries $j = 1, \ldots, J$ is weakly higher than revenue from the aggregate query $j = 0$. □

Proposition 3 shows that when the platform can design the auction for each query, finer information about what users want will never lower its expected revenue. This is because a platform can always ignore what it has learned and run the auction it used before. Thus, any change the platform makes leaves it at least as well off as before. Formally, Lemma 1 implies that expected revenue is the expected value of the highest adjusted bid net of the display cost. Each adjusted bid is a click-through rate multiplied by a term that depends only on the advertiser, so revenue responds to finer information in the same way welfare does in Proposition 1. Proposition 2 showed that a fixed auction can lose revenue when a refined query has a clear single winner. The optimal auction avoids this loss because its reserve, rather than the bid of a weak rival, sets the price. The winner pays at least the amount at which its adjusted bid covers the display cost. The result is closely related to Fu et al. (2012), who show that using data about user types raises revenue under the optimal auction. In our setting the information concerns queries and showing an ad is costly.

The finding implies that the effect of conversational search on advertising revenue depends less on the information itself than on whether the platform can adapt its auction to that information. A platform that keeps its reserves and weights fixed may earn less as queries become more specific, whereas one that adjusts them query by query does not face this decline. The optimal auction extracts more of advertisers' surplus and sometimes withholds ads that would create value. As a result, advertiser surplus and total welfare may not improve with more information.

3.3 Practical Implementation of Optimal Auctions

Major advertising platforms, including Google and Meta, sell ads through data-driven auctions with tunable parameters such as reserve prices and the weights that convert bids into rankings. Internal Google documents describe these parameters as "pricing knobs" that allow the platform to extract more of the value its auctions generate (Section 5).

Tuned auctions of this kind can be viewed as a practical, scalable way to approximate the optimal auction from bidding data. Proposition 4 makes this precise. When advertisers' demand elasticities are constant, the optimal auction for a single ad slot is a weighted second-price auction. This is also the format Google Search uses, with weights and a reserve chosen by the platform.

Proposition 4. *Assume that each $\epsilon_i(u_i)$ is constant in u_i. Then the optimal auction is a weighted second-price auction with reserve price $\rho = c$ and bidder weights*

$$w_{ij} = \alpha_{ij}\left(1 - \frac{1}{\epsilon_i}\right).$$

If, in addition, the elasticities are the same across bidders ($\epsilon_i = \epsilon > 1$), the optimal auction can be implemented with a weighted second-price auction with bidder weights equal to their click-through rates ($w_{ij} = \alpha_{ij}$) and a reserve price $\rho = \frac{c}{1-1/\epsilon}$.

The proof is in Appendix A.

Proposition 4 implies that the auction format Google Search already uses can be optimal. In a single-slot setting with constant demand elasticities, a weighted second-price auction maximizes revenue, provided that the weights (Google's "quality scores") reflect both an advertiser's click-through rate and the elasticity of its demand. Assuming a constant elasticity is a natural simplification as it lets the platform summarize each advertiser's demand with a single parameter that can be estimated from bidding data. If the platform goes further and assumes a common elasticity for all advertisers, the optimal weights reduce to the click-through rates themselves, and the platform's only remaining lever is the reserve price, which it sets above the display cost at $c/(1-1/\epsilon)$.

4 Timing of Ads in Conversations

In keyword search, ads appear alongside the results as soon as the user submits a query. In a conversation, this may not be the case. A user may start with a broad query and reveal more about what they are looking for over several exchanges. In response, the platform can choose when to show an ad, if at all. Showing an ad early ensures that the consumer will see an ad. Waiting lets the platform learn more about the user's intent and sell a better-targeted impression. For instance, a conversation that starts with "running shoes" may later turn to "Nike Vaporfly versus Adidas Adios Pro for a marathon under three hours," a request that is more specific and closer to the user's product and brand preferences. At the same time, waiting carries the risk that the consumer may leave the conversation before any ad is shown. This section studies that trade-off.

4.1 Setup

We keep the primitives of Section 2: advertiser i has value per click u_i drawn from F_i, and the platform pays a display cost c whenever it shows an ad. What is new is that the platform observes the conversation as it unfolds and can choose the turn at which it uses its advertising opportunity.

To study the trade-off between advertising early and late, we consider a conversation with at most two turns and a single advertising opportunity.[5] At turn 1, the platform observes the conversation history h_1 and either displays an ad or waits. If it shows the ad, the ad opportunity is used. If it waits, the user continues to turn 2 with probability $1-p$ and leaves with probability p. If the user continues, the platform can still show the ad at turn 2, after which the conversation ends. The platform decides when the auction is held; advertisers bid whenever it takes place and cannot choose the turn at which they compete. The timing problem is related to sequential screening, in which a seller contracts with a buyer who learns over time (Courty and Li, 2000).

In this setting, waiting is valuable for the platform because the next turn may reveal more about user preferences. Let θ denote the user's need and $\alpha_i(\theta)$ advertiser i's click-through rate for a user with need θ. The platform does not observe θ. At turn t, it evaluates each advertiser at its expected click-through rate given the conversation so far, $\alpha_{it} = \mathbb{E}[\alpha_i(\theta) \mid h_t]$. A second turn updates the expectations. Advertisers who

[5] We allow one ad per conversation since this is the simplest setting in which showing an ad early has an opportunity cost.

fit the revealed need become more attractive, and those who do not become less attractive. By waiting, the platform can therefore sell the single impression with more precise information about the user's intent.

Let $r \in \{\text{Fix}, \text{Flex}\}$ denote the auction regime, and let $R_t^r(h_t)$ be the platform's expected revenue, net of the display cost c, from holding the regime-r auction at turn t after history h_t. Advertisers' values per click are independent of the conversation, so the expectation is taken over these values, with click-through rates set to the platform's turn-t estimates α_{it}.

In the fixed regime, the platform uses a weighted second-price auction with weights equal to the current click-through rates and a reserve ρ that does not change across turns. Advertiser i's effective bid is $e_{it} = \alpha_{it} u_i$, and $e_{(1)t} \geq e_{(2)t}$ denote the two highest effective bids. The ad is shown if the highest effective bid clears the reserve, and the winner pays the larger of the second-highest effective bid and the reserve, so

$$R_t^{\text{Fix}}(h_t) = \mathbb{E}\left[\left(\max\{e_{(2)t}, \rho\} - c\right) \mathbf{1}\{e_{(1)t} \geq \rho\} \,\middle|\, h_t\right].$$

In the flexible regime, the platform runs the optimal auction of Lemma 1 for the current history. Its expected revenue is the expected value of the highest virtual value per impression net of the display cost, or zero if no ad is shown:

$$R_t^{\text{Flex}}(h_t) = \mathbb{E}\left[\max\left\{0,\, \alpha_{1t}\tilde{\psi}_1(u_1) - c,\, \ldots,\, \alpha_{It}\tilde{\psi}_I(u_I) - c\right\} \,\middle|\, h_t\right],$$

where $\tilde{\psi}_i$ is advertiser i's virtual value per click, defined in Section 3.

Revenue at each turn is thus given by the static analysis of Section 3, applied to the information available at that turn. The two turns are linked as follows. We say that the second turn *refines* the first if, for every advertiser i,

$$\mathbb{E}\left[\alpha_{i2} \mid h_1\right] = \alpha_{i1},$$

where the expectation is over the histories h_2 that can follow h_1.[6] Proposition 3 then implies that expected flexible-regime revenue at turn 2 is at least as high as at turn 1, while by Proposition 2, expected fixed-auction revenue may be higher or lower, because a more specific history can lower the second-highest effective bid.

The analysis helps clarify what can make advertising in the earlier stages of conversational search attractive. Under the flexible regime, waiting can only raise the expected revenue from the impression, so information alone never favors showing the ad early. The platform advertises at turn 1 only because waiting is costly: the user may leave, or an ad later in the conversation may be more disruptive. Under the fixed regime, information itself can favor showing the ad early. A more specific turn may reveal that one advertiser is by far the best match, leaving it with little competition and lowering the price it pays. A platform with a fixed auction may therefore prefer to show the ad early, while the user's intent is still uncertain, even when no user would leave the conversation. The selling rule thus affects the timing of conversational ads as well as their revenue. We next make this comparison precise by introducing the costs of waiting.

[6] Refinement holds by the law of iterated expectations whenever the user's decision to continue is unrelated to their need θ. Under refinement, the turn-1 history plays the role of the aggregate query in Section 3 and the possible turn-2 histories play the role of the granular queries.

4.2 Optimal Timing

Besides the display cost c, which is already included in $R_t^r(h_t)$, showing an ad at turn t imposes a user-experience cost $a_t \geq 0$ on the platform, for example through lower engagement or retention. This cost may depend on the turn, as interrupting a user who is still describing what they need may cost more than showing an ad after the consumer communicates their need.

At turn 2, the platform shows the ad if and only if doing so is profitable, so its value is $V_2^r(h_2) = \max\{R_2^r(h_2) - a_2, 0\}$. Let

$$W_1^r(h_1) := \mathbb{E}\left[\max\{R_2^r(h_2) - a_2, 0\} \,|\, h_1, \text{ no ad at turn 1}\right]$$

denote the expected value of reaching turn 2 with the opportunity still available. Showing the ad at turn 1 yields $R_1^r(h_1) - a_1$ and uses up the opportunity. Waiting preserves the opportunity, but a user reaches turn 2 only with probability $1 - p$. The value at turn 1 is therefore

$$V_1^r(h_1) = \max\left\{R_1^r(h_1) - a_1,\ (1-p)\, W_1^r(h_1)\right\}.$$

Proposition 5 (Optimal Stopping)**.** *In the two-turn conversation, the platform shows the ad at turn 1 if and only if*

$$R_1^r(h_1) - a_1 \geq (1-p)\, W_1^r(h_1).$$

It waits if and only if the reverse inequality holds.

The proof is in Appendix A.

Rewriting the condition in Proposition 5 as a threshold on the attrition probability gives the following corollary.

Corollary 1 (Continuation Risk)**.** *Suppose the distribution of h_2 conditional on reaching turn 2 does not depend on p. If $W_1^r(h_1) > 0$, waiting is optimal if and only if*

$$p \leq p^*(h_1) := 1 - \frac{R_1^r(h_1) - a_1}{W_1^r(h_1)}.$$

The proof is in Appendix A.

Proposition 5 compares selling an impression earlier with keeping the option to sell it later. The value of waiting depends on several factors. Users may leave, so the risk of attrition is a concern. Higher attrition makes early advertising more attractive, and a higher expected value of the second turn favors waiting.

Information can raise or lower the value of waiting, since a more specific second turn raises expected revenue under the flexible regime but may lower it under the fixed regime (Section 4.1). To isolate the role of information, consider the case without attrition ($p = 0$). If the second turn refines the earlier query, Proposition 3 implies that expected revenue under the flexible regime is at least as high at turn 2 as at turn 1, so waiting is weakly optimal whenever a later ad is no more disruptive ($a_2 \leq a_1$). With $p > 0$, the platform weighs this gain against the risk that the user leaves before an ad is shown.

4.3 Longer Conversations: Monetizing Sessions

The two-turn model extends directly to longer conversations. Suppose a conversation can last up to T turns, the platform still has one advertising opportunity, and the user continues from turn t to turn $t+1$ with probability $s_t(h_t)$ if no ad has been shown; in the two-turn model, $s_1 = 1 - p$. The platform's value at turn $t < T$ is

$$V_t^r(h_t) = \max\left\{R_t^r(h_t) - a_t,\ C_t^r(h_t)\right\}, \qquad C_t^r(h_t) := s_t(h_t)\,\mathbb{E}\left[V_{t+1}^r(h_{t+1}) \,\middle|\, h_t,\ \text{no ad at turn } t\right],$$

with $V_T^r(h_T) = \max\{R_T^r(h_T) - a_T,\, 0\}$.

Remark 1 (Longer Conversations)**.** By backward induction, the platform shows the ad at turn $t < T$ if and only if $R_t^r(h_t) - a_t \geq C_t^r(h_t)$, and at turn T if and only if $R_T^r(h_T) \geq a_T$.[7]

The rule has the same form as Proposition 5. At each turn, the platform compares the net revenue from advertising now with the value of keeping the opportunity, $C_t^r(h_t)$. With longer conversations, this value accumulates and each turn creates an opportunity for the consumer to reveal what they need or to leave. Because $C_t^r(h_t)$ increases with the continuation probabilities, waiting is more attractive in conversations where users tend to keep talking. And waiting is more attractive when showing an ad at the current turn is costly, for example because it would interrupt a user who is still describing their need. Conversely, the platform advertises early when users are likely to leave or when the current history already reveals what the user wants. Allowing the platform to choose when to advertise thus implies moving from selling an impression per query to selling an impression per conversation.

5 Discussion

The theoretical results show that richer query information has an ambiguous revenue effect when the selling rule is fixed, but raises expected revenue when the platform can adapt the mechanism to the realized query. This distinction is empirically relevant because modern advertising platforms no longer resemble a single sponsored-search auction with fixed parameters. Instead, they combine automated bidding, query expansion, prediction, and platform-controlled auction parameters. The tension between better matching and weaker competition has been studied directly for the sale of advertising impressions (Bergemann et al., 2021; Tunuguntla et al., 2026).

In the classic literature on sponsored-search auctions, advertisers compete in a single generalized second-price auction with fixed rules and parameters (Edelman et al., 2007; Varian, 2007). Today, sponsored search and related advertising markets, such as display and social media advertising, have evolved toward more complex systems (Google Ads Help, 2026a,b; Meta for Business, 2026a,b,c). They use automated bidding layers where advertisers delegate decision-making to the platform across large numbers of auctions (Bergemann et al., 2025). The auction layers themselves have flexible parameters, or "knobs," that are optimized based on data (as in our flexible auction regime).

[7] The same rule holds in an infinite-horizon version with bounded payoffs and a discount factor $\delta < 1$, in which the value function is the unique bounded solution of the corresponding Bellman equation; see Appendix A.

We therefore view the flexible auction regime as the most relevant benchmark for current sponsored search, for other major segments of digital advertising, and for the LLM advertising markets now being developed. The purpose of the institutional discussion is to show how the practical margins controlled by platforms map to the assumptions of the model.

Sponsored Search Market. Google holds a dominant position in internet search, with a worldwide search-engine market share of roughly 90% as of April 2026 (StatCounter Global Stats, 2026). Search advertising alone generated $114.2 billion in U.S. revenue in 2025, out of $294.6 billion in total internet advertising revenue (Interactive Advertising Bureau and PwC, 2026). Alphabet Inc. reported $224.5 billion in "Google Search & other" revenue in its 2025 Form 10-K, while Meta Platforms reported $196.2 billion in advertising revenue in 2025 (Alphabet Inc., 2026; Meta Platforms, Inc., 2026). Meta's advertising ecosystem shares similar features to Google's sponsored-search marketplace, including automated bidding, auction-based allocation, and flexible platform-controlled auction parameters.

Most advertisers purchase Google Search ads through Google Ads. Large advertisers use Google Search Ads 360, which offers more sophisticated tools to bid on Google Ads. Google Ads offers both search and non-search ads, such as YouTube and third-party display ads. For sponsored search, Google Ads runs keyword auctions, similar to the original model of Edelman et al. (2007), but with tunable parameters. Auction weights are based on "quality scores," which include expected click-through rates, and reserve prices are set based on data.

While advertisers can directly set bids, they are encouraged to choose an objective such as sales, leads, or website traffic, set a daily budget, and then hand over many operational decisions to "Smart Bidding" and "AI Max." AI Max lets advertisers bid on broad matches of keywords. For example, instead of bidding on "car insurance," an advertiser can broad-match to semantic queries like those in the introduction. AI Max can also auto-generate ad copy based on the query (Google Ads Help, 2026a,b).[8]

This delegation is central to the interpretation of our model. When advertisers specify objectives and constraints rather than bidding manually on every query, the platform has more room to use its private information about match quality and to adjust the auction environment around that information. Google Search is a useful case because public litigation records describe the relevant auction parameters directly.

Auction parameters on Google Search. In the Google Ads auction layer, individual impressions are still sold with a generalized second-price auction.[9] However, the auction parameters are optimized based on data. Two key parameters are the reserve prices and the quality-score weights that determine effective bids. Tuning these parameters is an old idea, going back at least to the influential paper by Lahaie and Pennock (2007). In a large field experiment at Yahoo!, Ostrovsky and Schwarz (2023) set reserve prices using optimal auction theory and found that revenue rose substantially, with the largest

[8] In the current Meta interface, the same delegation appears through Advantage+. Advertisers provide objectives, budgets, creative inputs, and some constraints. They can also choose campaign budgets, rely on highest-volume bidding, and accept automatic placements across Facebook, Instagram, Reels, Marketplace, Messenger, and Audience Network. They can allow creative variations to improve ad performance (Meta for Business, 2026a,b,c). Still, Meta controls the practical margins that govern ad matching, placement, creative delivery, and pricing.

[9] Google also has a vertically integrated stack to transact third-party display advertisements. These moved to a first-price auction in Google Ad Manager in 2019 (Google, 2019).

gains for keywords with few bidders.

The Google Search antitrust lawsuit revealed several of the auction parameters and how executives and engineers think about them (U.S. Department of Justice, 2023). The documents also describe executives' objectives in tuning the parameters. For example, a Google slide deck says:

> "Second Pricing works great most of the time, but there are failure scenarios." ... "We need a way to extract value more directly. We need pricing mechanisms with pricing knobs" (U.S. Department of Justice, 2023, trial exhibit UPX0889).

The deck presents squashing as the "canonical example." Squashing moves every bidder's predicted click-through rate toward the highest one in the auction, which raises the weights of bidders with low predicted click-through rates. According to the deck, auction pressure increases when the winner has the highest predicted click-through rate.

Another trial exhibit describes what executives see as the key knobs:

> "[We] also directly affect pricing through tunings of our auction mechanisms, in general through the three levers that are format pricing, squashing or reserves. We'll call this 'intentional' pricing." (U.S. Department of Justice, 2023, trial exhibit UPX0509)

Finally, the importance of tuning auction parameters based on data is discussed in another deck:

> "tuning in a nutshell: 1. measure things 2. do math 3. profit!!! [sic]," and that predicted click-through rates are "at best a consensual hallucination"

suggesting that quality scores are designed to increase auction revenue as in our Proposition 4 (U.S. Department of Justice, 2023, trial exhibit UPX0043).

Joint Information and Mechanism Design. A natural criticism of our results is the following. Consider the case where only the platform has knowledge of the click-through rates. Assume that the platform can jointly decide what information to give bidders and what mechanism to use (Bergemann and Pesendorfer, 2007; Eső and Szentes, 2007). Ex ante, bidders know only the click-through rates from the average aggregate query. With these assumptions, the platform can always choose to give bidders no information. Thus, by a trivial dominance argument, revenue is always weakly higher in the high-information regime.

This is a trivial but important point. But note that this result is weaker than Proposition 3. Proposition 3 shows that the platform does better even if bidders have access to the additional information. Moreover, our model shows that the platform chooses to use a particular form of the optimal auction, which as discussed above is qualitatively similar to auctions used by major platforms.

6 Conclusion

This paper studies how richer user information changes the revenue economics of advertising as we transition from keyword search to conversational search. We find that when the ad auction rule is

fixed, finer information weakly improves ad matching but can thin advertiser competition within realized auctions. As a result, the revenue effect of finer information in conversational search is ambiguous. When the platform can use a flexible auction mechanism and adapt reserves and weights to the realized query, the ambiguity disappears and richer information weakly raises expected platform revenue.

This finding has implications for how conversational search will change search platforms. There is much debate about how the move to LLMs may alter search. One relevant change is that conversational interfaces elicit longer and more contextual queries. Another is that advertising platforms increasingly control matching rules, bid adjustments, and reserve policies, all of which affect their ability to monetize ads. Our results imply that these two changes are complements, because richer query data are most valuable to a platform that can adjust its allocation and pricing rules.

The results regarding the timing of ads provide additional insights. Conversational systems have the flexibility of deciding when to show an ad. Delaying an ad carries the risk of missing the opportunity to show it if the consumer leaves. But waiting can also elicit more specific information from users and allow the platform to improve the matching of ads to users. Ads in conversational assistants are increasingly matched to the conversation as a whole rather than to a single query (OpenAI, 2026b). On the one hand, this suggests that answer-first interfaces may reduce the number of visible ad opportunities while increasing the value of some opportunities that survive. Consumers are sensitive to how many ads they are exposed to (Huang et al., 2018), so the reduction may improve their browsing experience as well. On the other hand, the growth of conversational search and the value of richer conversational data raise questions about platforms' incentives to elicit information from users. Our model suggests that, in conversational environments, platforms that can adapt their auctions have stronger incentives to elicit, retain, and monetize detailed user signals. In some settings, these signals may contain personal or sensitive information.[10] Consumers are sensitive to firms tracking their data (Johnson et al., 2020; Aridor et al., 2023), so eliciting information may impose privacy costs on some users. The information that makes conversational advertising more profitable may also make privacy concerns more acute. We do not model these costs, but they point to a tension between the interests of digital platforms and those of their users.

References

Saeed Alaei, Ali Makhdoumi, and Azarakhsh Malekian. Dynamic learning and optimal advertising mechanism for LLM platforms, 2026. URL https://ssrn.com/abstract=6212838. Working Paper, accepted at ACM EC 2026.

Alphabet Inc. Alphabet Inc. Form 10-K for fiscal year 2025. SEC filing, 2026. URL https://www.sec.gov/Archives/edgar/data/1652044/000165204426000018/goog-20251231.htm.

[10]There is some evidence that users share deeply personal information in conversational environments. OpenAI states that it "recognize[s] that people share personal and sensitive information with ChatGPT" (OpenAI, 2026c) and has described how ChatGPT handles sensitive conversations (OpenAI, 2025). Anthropic's privacy documentation addresses users who want to input sensitive data into Claude chats (Anthropic, 2026b). Google's Gemini privacy hub states that Gemini collects what users say and share, including files, videos, screens, and photos, and warns users not to enter confidential information (Google, 2026b).

Anthropic. Claude Is a Space to Think, February 2026a. URL https://www.anthropic.com/news/claude-is-a-space-to-think. Accessed April 28, 2026.

Anthropic. I Would Like to Input Sensitive Data into My Chats with Claude. Who Can View My Conversations? Claude Help Center, 2026b. URL https://support.claude.com/en/articles/8325621-i-would-like-to-input-sensitive-data-into-my-chats-with-claude-who-can-view-my-conversations. Updated May 22, 2026. Accessed October 20, 2026.

Alina Arefeva and Delong Meng. Revealing information in auctions: The optimal auction versus the second-price auction. *Economics Letters*, 204:109870, 2021.

Guy Aridor, Yeon-Koo Che, and Tobias Salz. The effect of privacy regulation on the data industry: Empirical evidence from GDPR. *RAND Journal of Economics*, 54(4):695–730, 2023.

Susan Athey and Glenn Ellison. Position auctions with consumer search. *Quarterly Journal of Economics*, 126(3):1213–1270, 2011.

Santiago Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Jieming Mao, Aranyak Mehta, Vahab Mirrokni, Renato Paes Leme, Di Wang, and Song Zuo. Position auctions in AI-generated content, 2025. URL https://arxiv.org/abs/2506.03309. arXiv:2506.03309.

Santiago R. Balseiro and Yonatan Gur. Learning in repeated auctions with budgets: Regret minimization and equilibrium. *Management Science*, 65(9):3952–3968, 2019.

Martino Banchio, Aranyak Mehta, and Andres Perlroth. Ads in conversations. arXiv:2403.11022, 2024.

Kendra Barnett and Paul Hiebert. Google's AI Search Era Is Here—But Its Ad Business Hangs in the Balance, June 2025. URL https://www.adweek.com/media/google-ai-search-ad-business-openai-perplexity/. ADWEEK. Originally published June 10, 2025; later updated for ADWEEK Magazine's August 5 issue. Accessed April 28, 2026.

Dirk Bergemann and Alessandro Bonatti. Targeting in advertising markets: Implications for offline versus online media. *RAND Journal of Economics*, 42(3):417–443, 2011.

Dirk Bergemann and Martin Pesendorfer. Information structures in optimal auctions. *Journal of Economic Theory*, 137(1):580–609, 2007.

Dirk Bergemann, Tibor Heumann, and Stephen Morris. Selling impressions: Efficiency vs. competition. Cowles Foundation Discussion Paper 2291, 2021.

Dirk Bergemann, Alessandro Bonatti, and Nicholas Wu. Bidding with budgets: Data-driven bid algorithms in digital advertising. *International Journal of Industrial Organization*, 102:103172, 2025.

Kshipra Bhawalkar, Alexandros Psomas, and Di Wang. Sponsored questions and how to auction them, 2025. URL https://arxiv.org/abs/2512.03975. arXiv:2512.03975.

Simon Board. Revealing information in auctions: The allocation effect. *Economic Theory*, 38(1):125–135, 2009.

Tim Bradshaw, Ramsay Hodgson, and Cristina Criddle. AI May Fatally Wound Web's Ad Model, Warns Tim Berners-Lee. Financial Times, November 2025. URL https://www.ft.com/content/20592619-1bb9-451e-b4af-fcd33d981076. Accessed August 12, 2026.

Athena Chapekis and Anna Lieb. Do People Click on Links in Google AI Summaries? Pew Research Center, July 2025. URL https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/. Accessed October 23, 2026.

Pascal Courty and Hao Li. Sequential screening. *Review of Economic Studies*, 67(4):697–717, 2000.

Francesco Decarolis and Gabriele Rovigatti. From Mad Men to Maths Men: Concentration and buyer power in online advertising. *American Economic Review*, 111(10):3299–3327, 2021.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. Auctions with LLM summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pages 713–722. ACM, 2024.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1): 242–259, 2007.

Péter Eső and Balázs Szentes. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies*, 74(3):705–731, 2007.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with LLMs: Opportunities and challenges. arXiv:2311.07601, 2023.

Ina Fried. Scoop: OpenAI Projects $100 Billion in Ad Revenue by 2030, April 2026. URL https://www.axios.com/2026/04/09/openai-100-billion-in-ad-revenue. Axios. Accessed April 28, 2026.

Hu Fu, Patrick R. Jordan, Mohammad Mahdian, Uri Nadav, Inbal Talgam-Cohen, and Sergei Vassilvitskii. Ad auctions with data. In *Algorithmic Game Theory*, volume 7615 of *Lecture Notes in Computer Science*, pages 168–179. Springer, 2012.

Juan-José Ganuza and José S. Penalva. Signal orderings based on dispersion and the supply of private information in auctions. *Econometrica*, 78(3):1007–1030, 2010.

Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

Google. Simplifying Programmatic: First Price Auctions for Google Ad Manager, March 2019. URL https://blog.google/products/admanager/simplifying-programmatic-first-price-auctions-google-ad-manager/. Accessed May 15, 2026.

Google. A new generation of ads for the AI era of search, May 2026a. URL https://blog.google/products/ads-commerce/google-marketing-live-search-ads/. Accessed October 21, 2026.

Google. Gemini Apps Privacy Hub. Gemini Apps Help, 2026b. URL https://support.google.com/gemini/answer/13594961?hl=en. Accessed October 20, 2026.

Google Ads Help. About Smart Bidding. Product documentation, 2026a. URL https://support.google.com/google-ads/answer/7065882. Accessed October 20, 2026.

Google Ads Help. How AI Max for search campaigns works. Product documentation, 2026b. URL https://support.google.com/google-ads/answer/15910187. Accessed October 20, 2026.

Brent D. Griffiths. Anthropic's Super Bowl Spot Skewers ChatGPT: "Ads Are Coming to AI, but Not to Claude", February 2026. URL https://www.businessinsider.com/anthropic-super-bowl-openai-chatgpt-ads-claude-2026-2. Business Insider. Accessed April 28, 2026.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. Ad auctions for LLMs via retrieval augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jason Huang, David H. Reiley, and Nickolai M. Riabov. Measuring consumer sensitivity to audio advertising: A field experiment on Pandora internet radio. Working Paper, 2018.

Patrick Hummel and R. Preston McAfee. When does improved targeting increase revenue? *ACM Transactions on Economics and Computation*, 5(1):4:1–4:19, 2016.

Interactive Advertising Bureau and PwC. Internet Advertising Revenue Report: Full-Year 2025 Results, April 2026. URL https://www.iab.com/wp-content/uploads/2026/04/IAB_PwC_Internet_Ad_Revenue_Report_Full_Year_2025_April_2026.pdf. Accessed May 15, 2026.

Garrett A. Johnson, Scott K. Shriver, and Shaoyin Du. Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1):33–51, 2020.

Peter Kafka. Why OpenAI's Sam Altman Thinks Ads on ChatGPT Will Be a Huge Business, March 2026. URL https://www.businessinsider.com/chatgpt-ads-openai-sam-altman-chart-paid-users-free-2026-3. Business Insider. Accessed April 28, 2026.

Sébastien Lahaie and David M. Pennock. Revenue analysis of a family of ranking rules for keyword auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce (EC '07)*, pages 50–56. ACM, 2007.

Poet Larsen and Davide Proserpio. How language models reshape sponsored search: Evidence from Google's BERT rollout. SSRN Working Paper 4614402, 2026.

Jonathan Levin and Paul Milgrom. Online advertising: Heterogeneity and conflation in market design. *American Economic Review: Papers and Proceedings*, 100(2):603–607, 2010.

LMSYS Organization. LMSYS-Chat-1M Dataset, 2023. URL https://huggingface.co/datasets/lmsys/lmsys-chat-1m. Hugging Face dataset repository. Accessed October 2026.

Aisha Malik. ChatGPT Reaches 900M Weekly Active Users. TechCrunch, February 2026. URL https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/. Accessed October 23, 2026.

Meta for Business. Advantage+ campaign budget. Product documentation, 2026a. URL https://www.facebook.com/business/ads/meta-advantage-plus/budget. Accessed September 2026.

Meta for Business. Advantage+ creative. Product documentation, 2026b. URL https://www.facebook.com/business/ads/meta-advantage-plus/creative. Accessed October 2026.

Meta for Business. Advantage+ placements. Product documentation, 2026c. URL https://www.facebook.com/business/ads/meta-advantage-plus/placements. Accessed September 2026.

Meta Platforms, Inc. Meta Platforms, Inc. Form 10-K for fiscal year 2025. SEC filing, 2026. URL https://www.sec.gov/Archives/edgar/data/1326801/000162828026003942/meta-20251231.htm.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Emma Nolan. A Google VP Explains Why Ads Make Sense in AI Search but Not Gemini—Yet, January 2026. URL https://www.businessinsider.com/google-vp-says-ads-arent-coming-to-gemini-yet-why-2026-1. Business Insider. Accessed April 28, 2026.

OpenAI. Strengthening ChatGPT's Responses in Sensitive Conversations, October 2025. URL https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/. Accessed October 20, 2026.

OpenAI. A milestone in expanding access to AI, August 2026a. URL https://openai.com/index/expanding-access-to-ai-with-chatgpt-ads/. Accessed October 21, 2026.

OpenAI. Ads in ChatGPT: The Basics. OpenAI Help Center, 2026b. URL https://help.openai.com/en/articles/20001207-ads-in-chatgpt-the-basics. Accessed October 23, 2026.

OpenAI. Introducing ChatGPT Health, January 2026c. URL https://openai.com/index/introducing-chatgpt-health/. Accessed October 20, 2026.

OpenAI. Testing Ads in ChatGPT, February 2026d. URL https://openai.com/index/testing-ads-in-chatgpt/. Accessed April 28, 2026.

Michael Ostrovsky and Michael Schwarz. Reserve prices in internet advertising auctions: A field experiment. *Journal of Political Economy*, 131(12):3352–3376, 2023.

Dominic Preston. Anthropic Says "Claude Will Remain Ad-Free," Unlike ChatGPT, February 2026. URL https://www.theverge.com/ai-artificial-intelligence/873686/anthropic-claude-ai-ad-free-super-bowl-advert-chatgpt. The Verge. Accessed April 28, 2026.

Reuters. OpenAI projects $2.5 billion in ad revenue this year, $100 billion by 2030, Axios reports, April 2026. URL https://www.reuters.com/business/media-telecom/openai-projects-25-billion-ad-revenue-this-year-100-billion-by-2030-axios-2026-04-09/.

Ermis Soumalias, Michael J. Curry, and Sven Seuken. Truthful aggregation of LLMs with an application to online advertising. arXiv:2405.05905, 2024.

StatCounter Global Stats. Search Engine Market Share Worldwide – April 2026, April 2026. URL https://gs.statcounter.com/search-engine-market-share?lang=en. Accessed May 2026.

Daniel Thomas. Publishers Race to Counter 'Google Zero' Threat as AI Changes Search Engines. Financial Times, August 2025. URL https://www.ft.com/content/f7a0eb8e-ff5b-42ae-b882-4815dbb38653. Accessed August 12, 2026.

Srinivas Tunuguntla, Carl F. Mela, and Jason Pratt. Targeting information in ad auction mechanisms, 2026. URL https://arxiv.org/abs/2601.09541.

U.S. Department of Justice. U.S. and Plaintiff States v. Google LLC [2020] – Trial Exhibits, September 2023. URL https://www.justice.gov/atr/us-and-plaintiff-states-v-google-llc-2020-trial-exhibits. Antitrust Division. Accessed May 14, 2026.

Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.

Chujie Zhao, Qun Hu, Shiping Song, Dagui Chen, Han Zhu, Jian Xu, and Bo Zheng. LLM-auction: Generative auction towards LLM-native advertising, 2025. URL https://arxiv.org/abs/2512.10551. arXiv:2512.10551.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *International Conference on Learning Representations (ICLR)*, 2024.

A Appendix: Proofs

A.1 Preliminaries

Definition 3 (Weighted Second-Price Auction)**.** Fix a query with CTR vector $\alpha = (\alpha_1, \ldots, \alpha_I)$, bidder weights $(w_1, \ldots, w_I)$, and reserve price ρ. Advertisers submit per-click bids $b = (b_1, \ldots, b_I)$, and bidder i's effective bid per impression is

$$e_i := b_i w_i.$$

Let $e_{(1)} \geq e_{(2)} \geq \cdots \geq e_{(I)}$ denote the order statistics of effective bids.

The mechanism operates as follows.

1. The impression is allocated to the advertiser with the highest effective bid, $e_{(1)}$, if and only if $e_{(1)} \geq \rho$. Otherwise, no ad is shown.

2. If the impression is allocated, the winner i^* pays per impression

$$\max(e_{(2)}, \rho).$$

   Equivalently, the winner's payment per click is

$$\frac{\max(e_{(2)}, \rho)}{w_{i^*}}.$$

A.2 Proofs

Proof of Proposition 2 for the case $I > 2$. For $I > 2$, it is enough to provide examples with $I = 3$, $c = 0$, and two equiprobable queries.

For a case in which segmentation raises revenue, let $u = (10, 10, 9)$ and

$$\alpha(\omega_1) = (1, 0, 1), \qquad \alpha(\omega_2) = (0, 1, 1).$$

In each query-specific auction, the second-highest value is 9, so expected revenue in the High-Information Regime is 9. In the aggregate auction,

$$\bar{\alpha} = (0.5, 0.5, 1),$$

which yields effective values $(5, 5, 9)$. Revenue in the Low-Information Regime is therefore 5.

For a case in which segmentation lowers revenue, let $u = (10, 10, 10)$ and

$$\alpha(\omega_1) = (1, 0, 0), \qquad \alpha(\omega_2) = (0, 1, 0).$$

In the High-Information Regime, each query-specific auction has only one bidder with positive value, so revenue is 0. In the aggregate auction,

$$\bar{\alpha} = (0.5, 0.5, 0),$$

which yields effective values $(5, 5, 0)$. Revenue in the Low-Information Regime is 5.

The effect of segmentation on revenue is therefore ambiguous for $I = 3$. The same conclusion extends immediately to any $I > 3$ by adding bidders with zero value. These examples use degenerate bidder-value distributions F_i, but the same results hold for proper distributions that are sufficiently close to these examples. □

Proof of Lemma 1. If $\alpha_{ij} = 0$, bidder i's value for the impression is zero, and its effective bid $\alpha_{ij}\tilde{\psi}_i(u_i) = 0$ never exceeds the display cost $c > 0$. Such a bidder never wins and can be dropped from the query-j auction. Assume therefore that $\alpha_{ij} > 0$.

The distribution of bidder i's value per impression $v_i = \alpha_{ij}u_i$ is

$$G_i(v_i) := F_i\left(\frac{v_i}{\alpha_{ij}}\right).$$

Myerson (1981) defines bidder i's virtual value as

$$\psi_i(v_i) := v_i - \frac{1 - G_i(v_i)}{G_i'(v_i)} = v_i - \alpha_{ij}\frac{1 - F_i(v_i/\alpha_{ij})}{f_i(v_i/\alpha_{ij})}.$$

Therefore,

$$\psi_i(v_i) = \alpha_{ij}\tilde{\psi}_i(u_i),$$

where $\tilde{\psi}$ is the virtual value function corresponding to the distribution of value per click:

$$\tilde{\psi}_i(u_i) := u_i - \frac{1 - F_i(u_i)}{f_i(u_i)} = \left(1 - \frac{1}{\epsilon_i(u_i)}\right)u_i.$$

Since $\alpha_{ij} > 0$, $\psi_i(v_i) = \alpha_{ij}\tilde{\psi}_i(v_i/\alpha_{ij})$ is strictly increasing in v_i because $\tilde{\psi}_i$ is, so each G_i is regular and the result follows from the Myerson (1981) optimal auction characterization. □

Proof of Proposition 4. From Lemma 1, the platform allocates the impression to the bidder with the highest effective bid, provided it exceeds the display cost c. The effective bid for bidder i on query j is given by

$$\alpha_{ij}\left(1 - \frac{1}{\epsilon_i}\right)u_i.$$

If we define the bidder weight as $w_{ij} = \alpha_{ij}\,(1 - 1/\epsilon_i)$, the effective bid is simply $w_{ij}u_i$. The optimal auction ranks bidders by $w_{ij}u_i$ and sets a minimum allocation threshold of $w_{ij}u_i \geq c$. This is exactly a weighted second-price auction with weights w_{ij} and a reserve price $\rho = c$ applied to the weighted bids.

Now, consider the case where elasticities are identical across all bidders, so $\epsilon_i = \epsilon$ for all i. The effective bid becomes

$$\alpha_{ij}\left(1 - \frac{1}{\epsilon}\right)u_i.$$

The term $(1 - 1/\epsilon)$ is a strictly positive constant. Because scaling all bids by a positive constant does not change their relative ranking, ordering bidders by their effective bids is mathematically equivalent to ordering them by $\alpha_{ij}u_i$. This corresponds to a weighted second-price auction where the weights are simply the click-through rates, $w_{ij} = \alpha_{ij}$.

To determine the equivalent reserve price ρ for these CTR-weighted bids, we apply the same scaling factor to the threshold condition. The platform allocates the impression if

$$\alpha_{ij}\left(1-\frac{1}{\epsilon}\right)u_i \geq c.$$

Dividing both sides by the constant factor yields the condition

$$\alpha_{ij}u_i \geq \frac{c}{1-\frac{1}{\epsilon}}.$$

Thus, the optimal auction can be implemented as a weighted second-price auction with weights $w_{ij}=\alpha_{ij}$ and a modified reserve price $\rho=\frac{c}{1-1/\epsilon}$. □

Proof of Proposition 5. At turn 2, the session ends after the platform acts. If the platform shows the ad, its payoff is $R_2^r(h_2)-a_2$. If it does not show the ad, its payoff is zero. Therefore

$$V_2^r(h_2)=\max\{R_2^r(h_2)-a_2,0\}.$$

At turn 1, showing the ad yields the platform $R_1^r(h_1)-a_1$ and uses up the advertising opportunity. Waiting gives no current advertising revenue and keeps the single impression available for turn 2. The user reaches turn 2 with probability $1-p$. Conditional on reaching turn 2, the platform obtains $V_2^r(h_2)$. Hence the value of waiting is

$$(1-p)\mathbb{E}\left[V_2^r(h_2)\mid h_1,\ \text{no ad at turn 1}\right].$$

The platform chooses the larger of the value of advertising immediately and the value of waiting. Substituting the expression for $V_2^r(h_2)$ gives the condition in Proposition 5. □

Proof of Corollary 1. By Proposition 5, waiting is optimal if and only if

$$(1-p)\mathbb{E}\left[\max\{R_2^r(h_2)-a_2,0\}\mid h_1,\ \text{no ad at turn 1}\right]\geq R_1^r(h_1)-a_1.$$

Using the definition of $W_1^r(h_1)$, this is

$$(1-p)W_1^r(h_1)\geq R_1^r(h_1)-a_1.$$

If $W_1^r(h_1)>0$, division by $W_1^r(h_1)$ gives

$$p\leq 1-\frac{R_1^r(h_1)-a_1}{W_1^r(h_1)}.$$

The comparative statics follow directly: increasing p lowers the left-hand side $(1-p)W_1^r(h_1)$, while increasing $W_1^r(h_1)$ raises it.

For the $p=0$ statement, if the second-turn state is a refinement of the first-turn state, the flexible-mechanism revenue result implies $\mathbb{E}[R_2^{\text{Flex}}(h_2)\mid h_1]\geq R_1^{\text{Flex}}(h_1)$ when the platform reaches turn 2. If

$a_2 \leq a_1$, then

$$\mathbb{E}\left[\max\{R_2^{\text{Flex}}(h_2) - a_2, 0\} \mid h_1\right] \geq R_1^{\text{Flex}}(h_1) - a_1,$$

so waiting weakly dominates early advertising. □

Proof of Remark 1. Consider first the finite-horizon case. The proof follows from the proof of Proposition 5, generalized to a longer time horizon. At the terminal turn T, there is no future advertising opportunity after the current decision. Showing the ad yields $R_T^r(h_T) - a_T$, while not showing the ad yields zero. Therefore

$$V_T^r(h_T) = \max\{R_T^r(h_T) - a_T, 0\},$$

and the platform shows the ad if and only if $R_T^r(h_T) - a_T \geq 0$.

Now fix any turn $t < T$ and suppose the continuation value V_{t+1}^r is given. If the platform shows the ad at h_t, it receives $R_t^r(h_t) - a_t$ and the advertising opportunity is used up. If it waits, it receives no current advertising revenue and reaches the next turn with probability $s_t(h_t)$. The value of waiting is therefore

$$C_t^r(h_t) = s_t(h_t)\mathbb{E}\left[V_{t+1}^r(h_{t+1}) \mid h_t, \text{ no ad at turn } t\right].$$

Hence

$$V_t^r(h_t) = \max\{R_t^r(h_t) - a_t, C_t^r(h_t)\},$$

and the platform shows the ad if and only if $R_t^r(h_t) - a_t \geq C_t^r(h_t)$. Backward induction from turn T gives the finite-horizon result.

For the infinite-horizon case, with bounded payoffs, we can define the Bellman operator

$$(\mathcal{T}V)(h) = \max\left\{R^r(h) - a(h), \delta s(h)\mathbb{E}[V(h') \mid h, \text{ no ad}]\right\}.$$

For any two bounded functions V and $\widetilde{V}$,

$$\|\mathcal{T}V - \mathcal{T}\widetilde{V}\|_\infty \leq \delta\|V - \widetilde{V}\|_\infty,$$

because $s(h) \leq 1$ and the maximum operator is nonexpansive. Since $\delta < 1$, $\mathcal{T}$ is a contraction and has a unique bounded fixed point. At that fixed point, the platform chooses the larger of current net revenue and discounted continuation value, giving the stated stopping rule.

□