Robust Learning-Aware Position Auctions for LLM-Estimated CTRs via LCB Optimization and Monotone Exploration

Liz Lemma Future Detective

January 16, 2026

Abstract

We study position auctions inside AI-generated content when click-through rates depend on both the advertiser and the generated position context, and must be estimated online using modern prediction systems (e.g., LLM-based relevance and fit models). Building on recent work that removes the separability assumption and reduces welfare/revenue maximization to a winner-determination problem (WDP) with externalities, we focus on the deployment reality that p_{ij} is not known and may drift or be biased. We propose a learning-aware mechanism that (i) uses lower confidence bounds (LCBs) on standalone CTRs to compute allocations by solving the WDP (for the MNL model) exactly on conservative estimates, and (ii) injects bid-independent randomization for exploration using a monotone bucketization rule. Because the resulting allocation is monotone in bids, payments can be computed by the envelope theorem (up to numerical discretization), yielding ε-truthfulness per round. We provide a regret-and-incentives theorem: under finite contexts and semi-bandit click feedback collected through single-ad exploration rounds, the mechanism achieves sublinear welfare regret relative to a clairvoyant benchmark while maintaining approximate dominant-strategy incentive compatibility. We also give robustness guarantees under bounded prediction bias, capturing manipulation or systematic model error common in LLM-era ad systems.

Table of Contents

3. 3. Model: contexts, standalone CTRs, MNL click model, information structure, feedback model (single-ad exploration providing direct Bernoulli samples), and benchmark definition (clairvoyant per-round welfare-optimal allocation).

4. 4. Mechanism: LCB-WDP allocation, monotone exploration (bucketization / single-edge sampling), and envelope-based payments with discretization; implementation details and computational complexity.

5. 5. Incentive analysis: monotonicity of the induced allocation; per-round ε-DSIC from payment discretization; discussion of what breaks if estimates depend on bids.

6. 6. Learning and regret: concentration of CTR estimates, decomposition of welfare regret into (i) estimation conservatism, (ii) exploration cost, and (iii) model-bias term; main regret theorem and corollaries for choosing γ_t and confidence widths.

7. 7. Robustness to manipulation and misspecification: bounded bias model; optional robust optimization variant (max-min over a confidence set) and when it requires numerical methods.

8. 8. Extensions (brief): cascade model with monotone bucketization; multi-objective constraints (trust budget); nonstationarity (sliding-window LCBs); empirical evaluation plan.

9. 9. Conclusion: implications for deployable assistant monetization and open problems (multi-click behavior, endogenous generation policies).

1 Introduction

Advertising inside conversational and generative interfaces is no longer well described by a fixed list of "slots" with stable, query-level click-through rates. In a 2026-style assistant, the interface itself is part of the outcome: the system chooses where an ad may appear (e.g., above the answer, inline as a cited recommendation, inside a tool call, or as a follow-up suggestion), how much surrounding text it displaces, and how it interacts with other rendered elements such as citations, images, and UI widgets. We refer to these candidate insertion points as *positions*, but the key modeling difference is that the set of positions and their salience are *context dependent*. The context includes the user query, the conversation state, and the UI template that will be rendered, and it changes from round to round.

This shift creates a new bottleneck: to allocate ads efficiently and fairly, the platform must learn the click response of users to *pairs* of (advertiser, position) within each context, while advertisers strategically choose bids in response to the allocation and pricing rule. Learning and incentives are tightly coupled. If we learn by experimenting with allocations that depend on bids, then the resulting feedback can be contaminated by strategic behavior and can invalidate standard concentration arguments. Conversely, if we ignore learning and treat predicted click probabilities as given, we inherit systematic miscalibration and bias that can distort welfare and, in practice, erode trust in both the platform and the auction.

We take the view that the right abstraction for these environments is a contextual choice model with externalities across displayed ads. The multinomial logit (MNL) family is a natural baseline: when multiple ads are shown, each ad competes for attention against other ads and against the outside option, and the click probability is governed by a simple and interpretable functional form. Importantly, MNL retains a tractable optimization structure: given per-ad "attractiveness" parameters (equivalently, log-odds), the welfare-maximizing slate can be computed by solving a winner determination problem (WDP) subject to a capacity constraint on the number of shown ads. In our setting, the MNL parameters are *contextual* and *unknown*, and must be learned online.

A central design goal is to preserve incentive properties while learning. In classical single-parameter auction design, dominant-strategy incentive compatibility (DSIC) is obtained by ensuring that the allocation rule is monotone in each bidder's report, and then charging the corresponding threshold-style payments via the envelope theorem. In contextual ad allocation, two obstacles arise. First, the allocation problem itself is combinatorial: we must choose a feasible matching between advertisers and positions up to capacity. Second, the allocation depends on estimated click probabilities, and if these estimates are updated using feedback generated under bid-dependent allocations, the monotonicity and the validity of the payment construction

become delicate.

Our approach isolates the learning system from strategic influence by separating *exploration* from *exploitation* in a way that is compatible with monotonicity. On exploration rounds, we show at most one ad, selected by a bid-monotone exploration rule, and we record a direct Bernoulli sample for a specific (context, advertiser, position) triple. This provides clean, interpretable feedback that can be aggregated into confidence intervals for the standalone click probability of that advertiser in that position under that context. On exploitation rounds, we commit to using only a conservative estimate of click probabilities—specifically, lower confidence bounds—and we compute the welfare-maximizing allocation for the MNL model under these bounds. This "optimism in reverse" is a deliberate choice: by underestimating click probabilities in a way that is independent of current bids, we can recover a monotone allocation rule and make truthful bidding approximately optimal, while still guaranteeing that learning progresses through dedicated, controlled experimentation.

The economic motivation for conservative bounds is straightforward. In an environment where advertisers pay per click and have private per-click values, overstating click probabilities effectively inflates the perceived marginal product of an advertiser's bid and can create incentives for bid shading or for gaming the prediction pipeline. Understating click probabilities, in contrast, may sacrifice some short-run welfare but yields a robust ranking signal for allocation and a stable payment map. This stability is particularly valuable in conversational interfaces, where small changes in UI placement can cause discontinuous changes in attention and where ex post explanations of outcomes (to advertisers, regulators, and users) often require monotone and auditable decision rules.

From an algorithmic standpoint, the MNL WDP serves as the backbone of the exploitation phase. Given context-dependent log-odds parameters (or equivalently, standalone click probabilities) for each advertiser-position pair, the platform chooses a feasible matching up to capacity that maximizes estimated welfare (bid times predicted click probability). The exact form of the MNL click probabilities captures substitution among displayed ads, making the optimization problem richer than independent-slot models yet still amenable to exact solving in the regime relevant to platform deployment. The resulting allocation rule is then combined with the exploration schedule to form a single randomized mechanism.

We summarize our contributions at three levels. First, we formalize a model of *generative positions* in which the platform must allocate advertisers to context-specific insertion points and users click according to an MNL choice model. The learning target is the standalone click probability for each (context, advertiser, position) triple, which is sufficient to parameterize the MNL probabilities under any feasible allocation. This separation between standalone and slate-level click behavior yields a modular estimation problem

and clarifies what must be learned to support welfare optimization.

Second, we provide a mechanism design template that couples confidence-bound learning with monotone allocation. The key ingredients are: (i) a conservative exploitation allocation obtained by solving the exact MNL WDP on lower confidence bounds; (ii) an exploration policy that samples at most one ad per exploration round to obtain unbiased Bernoulli feedback; and (iii) payments computed from the envelope formula applied to the realized allocation rule, using numerical integration with discretization step size η. Under this construction, the per-round allocation is monotone in each bid, and the mechanism is approximately DSIC with an explicit discretization-induced incentive loss of order $O(\bar{v}\,\eta)$. This is the practical price of computing payments in a complex allocation environment: we can make the incentive loss arbitrarily small at a controllable computational cost.

Third, we quantify the welfare consequences of learning. When contexts are drawn i.i.d. from a finite set and confidence intervals satisfy high-probability coverage uniformly over all (context, advertiser, position) triples, we obtain a sublinear regret guarantee relative to a clairvoyant benchmark that knows the true click parameters in each context. The resulting bound scales like $\tilde{O}\Big(\bar{v}\,\sqrt{|\mathcal{C}|\,nm\,T}\Big)$, plus an explicit exploration cost controlled by the exploration schedule. We also discuss robustness to systematic prediction bias: if the click estimator is adversarially miscalibrated by at most δ in expectation, then welfare regret degrades by an additive linear term $O(\bar{v}\,\delta T)$, making transparent the tradeoff between statistical learning error and structural model misspecification.

Beyond these technical statements, our broader message is that learning and incentives must be co-designed in generative advertising systems. The platform cannot treat click prediction as a purely statistical module, because the way it experiments changes the strategic environment; similarly, the mechanism cannot ignore the fact that click probabilities are learned, because payments and allocations built on unstable estimates are difficult to justify and easy to manipulate. By combining conservative confidence-bound optimization with a clean exploration channel and envelope-based pricing, we obtain a mechanism that is simultaneously implementable, approximately truthful, and equipped with explicit welfare guarantees. This provides a concrete bridge between WDP-based auction theory and the operational realities of context-rich, rapidly evolving conversational interfaces.

2 Related Work

Our setting sits at the intersection of three literatures that have largely evolved in parallel: (i) position auctions with *click externalities* and structured user choice models, (ii) truthful (or approximately truthful) mechanism design for *combinatorial* ad allocation problems where monotonicity is

the binding constraint, and (iii) online learning—bandits and learning-to-rank—in environments where the data-generating process is affected by the platform's allocation decisions and, in market settings, by bidders' strategic responses. A fourth strand concerns robustness to miscalibration and model misspecification, which becomes especially salient when click prediction is delegated to complex ML systems.

Position auctions beyond independent slots. The classical sponsored search abstraction assumes separable click-through rates: a position effect times an advertiser effect, with ads competing only through the platform's ranking rule. This foundation underlies the analysis of generalized second price and related designs (e.g., **??**). As soon as we move to generative interfaces—where "positions" are context-specific insertion points embedded in text, tool calls, or follow-ups—the independent-slot model becomes brittle. In particular, showing one ad can change the attention available to others and to the outside option. This motivates *externality* models of clicks, including cascade models (users scan ads sequentially and may stop after clicking) and discrete-choice models such as the multinomial logit (MNL), where each shown item competes with every other shown item and the no-click option.

There is a substantial literature analyzing welfare and revenue under cascade- and MNL-type click models in ad auctions and assortment problems (see, among many others, **???**). A useful conceptual takeaway from this work is that the platform's optimization problem is no longer a simple sort-by-score. Instead, it becomes a winner determination problem (WDP) over slates subject to feasibility constraints (e.g., each position used once, each advertiser at most once, and a total capacity constraint). Our mechanism leverages this slate structure directly: we treat the MNL model as the click response function during exploitation and solve the corresponding WDP exactly (given parameters), rather than imposing an approximation that restores separability.

Truthfulness, monotonicity, and approximation. In single-parameter environments, DSIC reduces to monotonicity of the allocation rule in each bidder's report, plus payments computed by the envelope formula. For position auctions with externalities, achieving monotonicity is technically delicate because the welfare-maximizing slate can change discontinuously with bids, and because externalities create complementarities and substitution effects that complicate standard greedy arguments. A significant body of work therefore focuses on truthful approximations for combinatorial allocation problems, where the primary design constraint is to ensure monotonicity while retaining computational tractability (e.g., **??**).

Within the externality click models, two approaches are common. One is to design truthful mechanisms for restricted structure (e.g., cascade mod-

els with specific scanning assumptions) or to impose monotone allocation heuristics that admit threshold payments. The other is to accept approximation in welfare (or revenue) in exchange for a monotone rule, sometimes via maximal-in-range constructions or carefully designed rounding schemes. Our work follows a different path: we retain the exact MNL WDP as the exploitation backbone, but we modify the *inputs* to the WDP by replacing unknown click parameters with bid-independent lower confidence bounds. This "conservative WDP" perspective is, for us, the key to reconciling exact slate optimization with incentive constraints in a learning environment. It also clarifies what is, and is not, being approximated: the optimization is exact for the surrogate parameters, while statistical uncertainty is handled via confidence intervals and an explicit exploration channel.

It is also useful to contrast our approach with monotone approximations developed for cascade-style models. In that literature, one often obtains monotonicity by restricting attention to allocation rules that preserve an ordered structure (e.g., placing ads in a fixed sequence and deciding which to include), which can yield tractable threshold payments but may mismatch the interaction patterns of generative UI placements. The monotone cascade approximation is attractive when it is a faithful behavioral model; our contribution is to show that, when the platform commits to an MNL choice model and is willing to solve the corresponding WDP, one can still preserve (approximate) truthfulness by using conservative parameter estimates and by isolating learning from bids.

Online learning in auctions and learning-to-rank. The learning problem in ad allocation is often framed as a multi-armed bandit or contextual bandit, where the platform must trade off exploration (to learn CTRs) against exploitation (to maximize immediate welfare or revenue). This viewpoint underlies much of the online advertising and recommendation literature, including work on learning-to-rank with click feedback and click models tailored to ranked lists (e.g., **???**). However, these algorithms typically take the ranking objective as given and do not treat advertisers as strategic agents whose bids respond to the allocation and pricing rule.

A smaller but growing literature studies bandit learning under incentive constraints, where the platform must learn unknown parameters (such as CTRs) while maintaining truthful or approximately truthful bidding (e.g., **???**). Two recurring lessons from this line of work are central to our design. First, if the data used for learning is collected under bid-dependent allocations, then strategic bidders can influence the estimation pipeline, undermining both statistical guarantees and incentive properties. Second, truthful learning mechanisms often require an explicit separation between exploration and exploitation, or at least a careful accounting that ensures the learning rule is not manipulable by individual bids.

Our mechanism follows this separation principle in a particularly stark form: exploration rounds show at most one ad, generating a direct Bernoulli sample for a specific advertiser-position-context triple. This design choice sacrifices some immediate welfare but yields clean identification of *standalone* click probabilities and makes the estimation update bid-independent by construction. In turn, bid-independent confidence bounds can be treated as fixed inputs when establishing monotonicity of the exploitation allocation and when applying the envelope formula for payments.

Robustness, miscalibration, and ML prediction systems. Finally, there is a broad literature on robust mechanism design and on auctions with uncertainty about key primitives (e.g., **??**). In practice, click probabilities are outputs of complex prediction pipelines, potentially subject to systematic bias due to distribution shift, interface changes, or strategic feedback loops. From an economic perspective, misspecified CTRs are not merely "noise": they can change the platform's effective objective and can create incentives for advertisers to redirect effort toward gaming measurement rather than improving product quality.

Our analysis explicitly separates statistical uncertainty (handled via confidence intervals and exploration) from systematic bias (captured by an additive robustness parameter). This is complementary to robust design approaches that optimize worst-case objectives or impose ex post constraints. The main message is operational: even when one cannot guarantee perfect calibration, it is still valuable to (i) preserve a bid-independent learning channel, so that errors do not become strategically amplified, and (ii) use conservative estimates in the allocation rule, so that the mechanism does not over-react to optimistic predictions. At the same time, we acknowledge a limitation: conservative bounds can be welfare-reducing in the short run, and bounded-bias guarantees are only as meaningful as the monitoring and auditing procedures that justify a bound. This connects directly to practice and policy: transparency requirements and external audits of ad delivery and measurement can be interpreted as institutional mechanisms for shrinking the bias parameter and, therefore, the long-run welfare loss.

Positioning. Relative to the existing work on externality click models, we view our contribution as a synthesis tailored to generative positions: we keep the expressive MNL interaction structure and its exact WDP, but we couple it with a learning rule and a payment construction that preserve monotonicity despite unknown, context-dependent click parameters. Relative to truthful learning-in-auctions work, our novelty is in treating the *advertiser-position-context* triple as the estimation target and in emphasizing a clean exploration design that yields direct Bernoulli samples, which lets us state regret guarantees against a clairvoyant MNL benchmark while maintaining

approximate DSIC through envelope pricing with explicit discretization error.

3 Model

We model a repeated allocation problem in which a platform must select, in each interaction, a small slate of sponsored items to insert into a context-dependent interface. The central economic friction is that click response is both *contextual* and subject to *externalities* across displayed ads, while the platform must learn these response parameters online from click feedback that is itself shaped by the platform's past allocations. At the same time, advertisers are strategic and submit bids each round.

Rounds, contexts, and feasible slates. Time is discrete with rounds indexed by $t \in \{1, \dots, T\}$. At the start of round t, Nature draws a context c_t from a finite set $\mathcal{C}$, and the context is observed by both the platform and all advertisers. We interpret c_t broadly as the realized user query and conversation state together with UI features (e.g., available insertion points, layout, and surface), since these jointly determine the set of plausible ad placements and the user attention environment.

There are n advertisers (agents) and m candidate positions (insertion points) that may be available in a given context. The platform may allocate at most $K \leq m$ ads per round. We represent an allocation as a partial matching between advertisers and positions, encoded by an indicator matrix $\mathbf{x}_t = (x_{ij,t}) \in \{0,1\}^{n \times m}$ with feasibility constraints: each advertiser appears at most once, each position is used at most once, and the total number of matches is at most K. We write $\mathcal{X}(K)$ for this feasible set. Thus, $\mathbf{x}_t \in \mathcal{X}(K)$ captures both *which* advertisers are shown and *where* they are inserted.

Advertiser values and bids. Advertiser i has a private per-click value $v_i \in [0, \bar{v}]$ (single-parameter, quasi-linear). In each round t, advertiser i submits a nonnegative bid (report) $b_{i,t} \geq 0$, forming the bid vector $\mathbf{b}_t = (b_{1,t}, \dots, b_{n,t})$. We interpret $b_{i,t}$ as a value per click, so that welfare and transfers can be written naturally in per-click units. We allow bids to vary across rounds to accommodate dynamic bidding behavior; our incentive analysis will be per-round with respect to the induced allocation rule in that round.

Standalone click probabilities as primitives. The fundamental unknown primitives are *standalone* click probabilities. For each context $c \in \mathcal{C}$, advertiser i, and position j, we define

$$p_{ij}^{c} \in (0,1)$$

as the probability that a user clicks advertiser i when i is shown *alone* in position j in context c (i.e., no other ads are shown). These probabilities capture the joint effect of relevance, creative quality, and position-specific visibility in that context, abstracting from competitive interactions with other displayed ads. The matrix $p^c = (p^c_{ij}) \in (0,1)^{n \times m}$ is unknown to the platform.

We emphasize the operational interpretation: standalone probabilities are identifiable from randomized single-ad displays, and they can be estimated without modeling strategic interactions among advertisers. This motivates treating (c, i, j) as the estimation unit throughout.

From standalone probabilities to externalities: an MNL click model. When multiple ads are shown simultaneously, we assume user choice follows a multinomial logit (MNL) model. For each context c, advertiser i, and position j, we map the standalone probability to an MNL "utility" (log-odds)

$$\rho^c_{ij} \;=\; \log\left(\frac{p^c_{ij}}{1 - p^c_{ij}}\right).$$

The outside option (no click) is normalized to utility 0. Given an allocation $\mathbf{x}_t \in \mathcal{X}(K)$ in context c_t, the probability that the user clicks advertiser i at round t is

$$\pi_{i,t}(\mathbf{x}_t; \rho^{c_t}) \;=\; \frac{\sum_{j=1}^{m} x_{ij,t} \exp(\rho^{c_t}_{ij})}{1 + \sum_{i'=1}^{n} \sum_{j=1}^{m} x_{i'j,t} \exp(\rho^{c_t}_{i'j})},$$

and the no-click probability is the remaining mass,

$$\pi_{0,t}(\mathbf{x}_t; \rho^{c_t}) \;=\; \frac{1}{1 + \sum_{i'=1}^{n} \sum_{j=1}^{m} x_{i'j,t} \exp(\rho^{c_t}_{i'j})}.$$

Because each advertiser can be assigned to at most one position, the numerator for $\pi_{i,t}$ selects the (at most one) assigned position for i. The externality is immediate: increasing the attractiveness of one displayed ad raises the denominator and reduces the click probabilities of other displayed ads and of the outside option. This captures the "competition for attention" that is natural in generative interfaces where multiple insertions share a limited interaction budget.

The particular log-odds parameterization is convenient for two reasons. First, it preserves the interpretation of p^c_{ij} as a standalone click probability when $K = 1$: if exactly one ad is shown, say $x_{ij,t} = 1$ and no other matches are selected, then $\pi_{i,t} = p^{c_t}_{ij}$. Second, it permits a compact expression for welfare and comparative statics as functions of exponentiated parameters.

Platform objective and welfare in a round. Given true values $\mathbf{v}$ and an allocation $\mathbf{x}_t$, the (true) expected welfare in round t is the expected value

of the clicked advertiser,

$$\mathrm{W}_t(\mathbf{x}_t) \;=\; \sum_{i=1}^{n} v_i \, \pi_{i,t}(\mathbf{x}_t; \rho^{c_t}).$$

We treat welfare as the main performance criterion because it aligns with allocative efficiency and provides a clean benchmark for learning; later, the mechanism uses bids as proxies for values, with payments ensuring (approximate) incentive alignment.

Information structure and bid-independence of learning. A key modeling assumption is the separation between what advertisers can influence through bidding and what the estimation system uses for learning. The mechanism, tie-breaking rules, exploration schedule, and the MNL functional form are common knowledge. Values v_i are private information. The platform observes $(c_t, \mathbf{b}_t)$ each round, chooses an allocation, and observes click feedback. Critically, the platform maintains estimators $\hat{p}^c_{ij,t}$ and associated confidence intervals $[\underline{p}^c_{ij,t}, \overline{p}^c_{ij,t}]$ using *exploration data only*, where exploration is defined below. Because exploration allocations are constructed to be bid-independent (and to generate direct samples for a specified triple (c, i, j)), the resulting confidence bounds are statistically independent of current bids. This bid-independence is the linchpin that later allows us to treat the learning state as fixed when analyzing monotonicity of the allocation rule in bids.

Feedback model and single-ad exploration samples. We distinguish two types of rounds from the perspective of learning. In an *exploration* round, the platform shows at most one ad: it selects a pair (i, j) and sets $x_{ij,t} = 1$ with all other entries zero. The click feedback is then a Bernoulli random variable with mean $p^{c_t}_{ij}$, providing a direct sample for that advertiser–position–context triple. We denote by $N^c_{ij,t}$ the number of exploration samples collected for (c, i, j) up to time t. This sampling scheme deliberately avoids confounding from competitive effects and avoids the need to infer per-ad click propensities from slate-level outcomes under externalities.

In *non-exploration* rounds (exploitation rounds), the platform may show up to K ads and the realized click outcome is generated according to the MNL choice probabilities above. Depending on the application, the platform may observe the identity of the clicked ad (or simply whether a click occurred). Our cleanest learning guarantees do not require using exploitation feedback for estimation; it can be incorporated in practice, but doing so introduces additional modeling assumptions about observation noise and counterfactual inference under externalities. Accordingly, we treat exploitation feedback as optional and focus on exploration-only updates to $\hat{p}$ and its confidence bounds.

Benchmark: clairvoyant per-round optimal welfare. To evaluate learning performance, we compare the platform's achieved welfare to a clairvoyant benchmark that knows the true standalone click matrix for the realized context. In each round t, define the *clairvoyant* optimal welfare as

$$\mathrm{W}_t^* \;=\; \max_{\mathbf{x}\in\mathcal{X}(K)} \sum_{i=1}^{n} v_i\, \pi_{i,t}(\mathbf{x};\rho^{c_t}),$$

where ρ^{c_t} is computed from the true p^{c_t}. This benchmark is per-round and context-dependent; it corresponds to the welfare-maximizing MNL slate under the platform's feasibility constraints. Our regret metric aggregates the difference between this benchmark and the welfare achieved by the mechanism:

$$\mathrm{Reg}(T) \;=\; \sum_{t=1}^{T} \big(\mathrm{W}_t^* - \mathrm{W}_t\big).$$

This definition isolates the statistical and incentive frictions of interest: regret arises because the platform does not initially know p_{ij}^{c} and must learn it from exploration, and because the allocation and payment rules must be chosen to manage strategic bidding while operating under externalities.

The next section specifies the mechanism that couples (i) an allocation rule based on conservative estimates of p_{ij}^{c}, (ii) an explicit, monotone exploration policy that generates the direct Bernoulli samples above, and (iii) payments computed from the induced monotone allocation rule.

4 Mechanism

In each round, the platform couples a conservative exploitation rule based on *lower* confidence bounds with an explicit single-ad exploration policy that generates statistically clean samples. Payments are computed from the induced (randomized) monotone allocation rule via the envelope formula, implemented with a discretization step. The design goal is to separate (i) statistical learning of standalone click primitives from (ii) strategic bidding under externalities, while keeping the per-round optimization computationally tractable.

Step 0: confidence bounds and log-odds. At the start of round t, after observing context c_t and bids $\mathbf{b}_t$, the platform treats its learning state as fixed and computes lower confidence bounds $\underline{p}_{ij,t}^{c_t}$ for all advertiser–position pairs. We work in log-odds form,

$$\underline{\rho}_{ij,t}^{c_t} \;=\; \log\!\left(\frac{\underline{p}_{ij,t}^{c_t}}{1-\underline{p}_{ij,t}^{c_t}}\right),$$

and define the associated attractiveness weights

$$a_{ij,t} \;=\; \exp(\underline{\rho}^{c_t}_{ij,t}) \;=\; \frac{\underline{p}^{c_t}_{ij,t}}{1-\underline{p}^{c_t}_{ij,t}}.$$

These quantities are computed solely from exploration data and are therefore statistically (and strategically) independent of the current bid vector.

Exploitation allocation: LCB winner determination under MNL. With probability $1-\gamma_t$, the platform runs exploitation. Given bids $\mathbf{b}_t$, it chooses a feasible matching $\mathbf{x}_t \in \mathcal{X}(K)$ that maximizes bid-weighted click welfare under the MNL model using the conservative parameters $\underline{\rho}^{c_t}_t$:

$$\mathbf{x}_t \in \arg\max_{\mathbf{x}\in\mathcal{X}(K)} \sum_{i=1}^{n} b_{i,t}\,\pi_{i,t}(\mathbf{x};\underline{\rho}^{c_t}_t).$$

Using the fact that $\pi_{i,t}$ depends on $\mathbf{x}$ only through the selected edges, we can rewrite the objective as a ratio. For a matching $\mathbf{x}$, let

$$A_t(\mathbf{x}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\,a_{ij,t}, \qquad B_t(\mathbf{x}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\,b_{i,t}\,a_{ij,t},$$

so that under $\underline{\rho}$ the total expected *bid-weighted* welfare is

$$\sum_i b_{i,t}\pi_{i,t}(\mathbf{x};\underline{\rho}^{c_t}_t) \;=\; \frac{B_t(\mathbf{x})}{1+A_t(\mathbf{x})}.$$

Hence exploitation solves a fractional matching problem,

$$\max_{\mathbf{x}\in\mathcal{X}(K)} \frac{B_t(\mathbf{x})}{1+A_t(\mathbf{x})},$$

with deterministic tie-breaking (fixed in advance) to ensure a well-defined allocation rule.

Implementing the exploitation WDP. Although the objective is fractional, it admits standard reductions. One convenient approach is Dinkelbach's transform: for a scalar $\lambda \geq 0$, define

$$F_t(\mathbf{x};\lambda) \;=\; B_t(\mathbf{x}) - \lambda\big(1+A_t(\mathbf{x})\big) \;=\; \sum_{i,j} x_{ij}\,(b_{i,t}-\lambda)\,a_{ij,t} \;-\; \lambda.$$

For fixed λ, maximizing $F_t(\mathbf{x};\lambda)$ over $\mathcal{X}(K)$ is a maximum-weight matching problem with edge weights

$$w_{ij,t}(\lambda) \;=\; (b_{i,t}-\lambda)\,a_{ij,t},$$

since the constant $-\lambda$ does not affect the argmax. Thus, each inner step can be solved by an assignment solver (e.g., Hungarian algorithm after adding dummy advertisers/positions to model capacity K). Dinkelbach's iterations update λ to the achieved ratio $B_t(\mathbf{x})/(1 + A_t(\mathbf{x}))$ and converge to the optimal fractional value; because the feasible set is finite, this yields an exact optimum up to numerical tolerance. In what follows we treat the winner determination problem (WDP) as a black-box routine that returns an exact maximizer under $\underline{\rho}_t^{c_t}$.

Exploration allocation: single-edge sampling with bid-independence. With probability γ_t, the platform explores by showing *at most one ad*. Concretely, it selects a pair (i_t, j_t) as a function of (c_t, t) and the current exploration counts $\{N_{ij,t}^{c_t}\}_{i,j}$, and sets $x_{i_t j_t, t} = 1$ with all other entries zero. The click feedback is then a Bernoulli draw with mean $p_{i_t j_t}^{c_t}$, which is stored as a direct sample for the triple (c_t, i_t, j_t).

The selection rule for (i_t, j_t) can be implemented in several equivalent ways, all chosen to satisfy two requirements: (i) *coverage*: every (c, i, j) is sampled often enough (as a function of the number of times context c appears) to shrink confidence intervals; and (ii) *bid-independence*: the distribution of (i_t, j_t) is independent of $\mathbf{b}_t$ so that the resulting samples are not manipulable by current bids. A simple instantiation is "least-sampled" exploration: in context c_t, pick $(i_t, j_t) \in \arg\min_{i,j} N_{ij,t}^{c_t}$ with deterministic tie-breaking. More generally, one may *bucketize* the space of edges within each context into groups and cycle through buckets to smooth coverage across advertisers and positions; the key property is that the mapping from $(c_t, t, \{N_{ij,t}^{c_t}\})$ to (i_t, j_t) is fixed ex ante and does not depend on bids.

Because exploration shows a single ad, it also avoids having to invert MNL externalities to obtain per-edge estimates: the observed click is a clean sample of the standalone primitive, which is precisely the object entering the confidence bounds used by exploitation.

Payments via the envelope formula and numerical discretization. Fix a round t and view the platform's randomization (the exploration coin and any internal randomness in exploration edge selection) as realized after bids are submitted. For each realized random seed, the mechanism induces a deterministic allocation rule, and hence an interim click-through probability for advertiser i,

$$y_{i,t}(\mathbf{b}_t) \;=\; \pi_{i,t}(\mathbf{x}_t(\mathbf{b}_t); \underline{\rho}_t^{c_t}),$$

where $\mathbf{x}_t(\mathbf{b}_t)$ is the realized allocation (either the exploitation matching or the single-edge exploration display). Note that once $\mathbf{x}_t$ is fixed, $y_{i,t}$ depends on $\underline{\rho}$ and the MNL denominator, but *not* directly on $b_{i,t}$; bids influence $y_{i,t}$ only through which allocation is selected.

Given a monotone allocation rule (established in the next section), we compute payments using the standard single-parameter envelope formula:

$$t_{i,t}(\mathbf{b}_t) \;=\; b_{i,t}\, y_{i,t}(\mathbf{b}_t) \;-\; \int_0^{b_{i,t}} y_{i,t}\big(z, \mathbf{b}_{-i,t}\big)\, dz,$$

with the convention that non-displayed advertisers (those with $y_{i,t} = 0$) pay zero. In exploration realizations, $y_{i,t}$ is bid-independent, so the formula yields (up to numerical error) zero payment; exploitation realizations generate positive payments when an advertiser's bid affects selection into the slate.

To avoid computing the integral exactly, we approximate it on a grid with step size $\eta > 0$. Let $G(b_{i,t}) = \{0, \eta, 2\eta, \ldots, \lfloor b_{i,t}/\eta \rfloor \eta\}$ and define the left Riemann sum

$$\widehat{I}_{i,t}(\mathbf{b}_t) \;=\; \eta \sum_{z \in G(b_{i,t})} y_{i,t}\big(z, \mathbf{b}_{-i,t}\big).$$

We then charge

$$\widehat{t}_{i,t}(\mathbf{b}_t) \;=\; b_{i,t}\, y_{i,t}(\mathbf{b}_t) \;-\; \widehat{I}_{i,t}(\mathbf{b}_t).$$

Since $y_{i,t} \in [0,1]$ and bids are bounded by $\bar{v}$ (or capped at $\bar{v}$ without loss for welfare), this discretization introduces an additive per-round error that scales linearly with η, which we summarize as $\varepsilon_{IC}(\eta) = O(\bar{v}\,\eta)$ in the incentive discussion.

Computational complexity and practical remarks. Per round, the dominant cost is solving the exploitation WDP. Using Dinkelbach's method, each iteration requires one maximum-weight matching solve with weights $w_{ij,t}(\lambda)$; with a Hungarian-style algorithm this is polynomial in $n+m$ (after padding with dummies to encode capacity K). The number of Dinkelbach iterations needed for a target numerical tolerance is typically modest in practice and can be treated as logarithmic in the inverse tolerance.

Payment computation is potentially more expensive if implemented naïvely, because $\widehat{I}_{i,t}$ requires evaluating $y_{i,t}(z, \mathbf{b}_{-i,t})$ at $O(\bar{v}/\eta)$ grid points, and each evaluation entails re-solving the WDP with bidder i's bid replaced by z. Two standard mitigations are available: (i) compute payments only for advertisers who are allocated positive click probability in the realized allocation (others pay zero), and (ii) exploit monotonicity to replace a full grid sweep by a coarser adaptive grid or a search over critical bid thresholds. We keep the discretized envelope as the canonical implementation because it is transparent and yields a direct η–accuracy parameter that will map cleanly into the per-round incentive approximation guarantee.

Finally, we emphasize that exploration updates are performed *only* from single-ad exploration rounds. This choice is conservative but crucial: it

makes the confidence sequences and hence the exploitation objective statistically independent of current bids, which is exactly the separation needed for the monotonicity and incentive arguments that follow.

5 Incentive analysis

Our incentive claims rest on a clean separation principle: within a fixed round t, the learning state—hence the attractiveness weights $a_{ij,t}$ (equivalently, the log-odds matrix $\underline{\rho}_t^{c_t}$)—is treated as fixed and bid-independent, and the only strategic input is the bid vector $\mathbf{b}_t$. Under this separation, the round-t allocation is a single-parameter maximization of a weighted objective and therefore satisfies the monotonicity property required by the envelope formula.

Click probability as a quantity function. Fix round t and suppress the t subscript for readability. For any feasible matching $\mathbf{x} \in \mathcal{X}(K)$ and fixed attractiveness weights $a_{ij} > 0$, the MNL model implied by $\underline{\rho}$ induces for each advertiser i an interim click-through quantity

$$q_i(\mathbf{x}) \;=\; \pi_i(\mathbf{x}; \underline{\rho}) \;=\; \frac{\sum_{j=1}^{m} x_{ij} a_{ij}}{1 + \sum_{\ell=1}^{n} \sum_{k=1}^{m} x_{\ell k} a_{\ell k}},$$

with the convention that $q_i(\mathbf{x}) = 0$ if i is unmatched. Importantly, $q_i(\mathbf{x})$ depends on bids only through the selected matching $\mathbf{x}$; for fixed (c_t, t) and fixed confidence bounds, it is a deterministic function of $\mathbf{x}$ alone.

Monotonicity of the exploitation rule. In exploitation, the platform selects a matching $\mathbf{x}(\mathbf{b})$ maximizing the bid-weighted welfare $\sum_i b_i q_i(\mathbf{x})$, with a deterministic tie-breaking rule fixed ex ante. This is formally identical to the standard single-parameter allocation template "choose arg max of a weighted sum of quantities" once we interpret $q_i(\mathbf{x})$ as the quantity assigned to agent i.

Lemma 5.1 (Bid monotonicity in exploitation)**.** *Fix a round and learning state (hence fixed a_{ij}) and fix $\mathbf{b}_{-i}$. Let $\mathbf{x}(\mathbf{b})$ be the exploitation matching selected by the mechanism, and define $y_i(\mathbf{b}) = q_i(\mathbf{x}(\mathbf{b}))$. Then $y_i(b_i, \mathbf{b}_{-i})$ is weakly nondecreasing in b_i.*

Proof. Let $b_i < b_i'$ and write $\mathbf{b} = (b_i, \mathbf{b}_{-i})$ and $\mathbf{b}' = (b_i', \mathbf{b}_{-i})$. Let $\mathbf{x} = \mathbf{x}(\mathbf{b})$ and $\mathbf{x}' = \mathbf{x}(\mathbf{b}')$ denote the selected matchings (with deterministic tie-breaking). Optimality of $\mathbf{x}$ at $\mathbf{b}$ gives

$$b_i q_i(\mathbf{x}) + \sum_{k \neq i} b_k q_k(\mathbf{x}) \;\geq\; b_i q_i(\mathbf{x}') + \sum_{k \neq i} b_k q_k(\mathbf{x}').$$

Optimality of $\mathbf{x}'$ at $\mathbf{b}'$ gives

$$b'_i q_i(\mathbf{x}') + \sum_{k \neq i} b_k q_k(\mathbf{x}') \;\geq\; b'_i q_i(\mathbf{x}) + \sum_{k \neq i} b_k q_k(\mathbf{x}).$$

Adding these inequalities and canceling the common terms yields

$$(b'_i - b_i)\big(q_i(\mathbf{x}') - q_i(\mathbf{x})\big) \;\geq\; 0,$$

and since $b'_i - b_i > 0$ we conclude $q_i(\mathbf{x}') \geq q_i(\mathbf{x})$, i.e., $y_i(b'_i, \mathbf{b}_{-i}) \geq y_i(b_i, \mathbf{b}_{-i})$. □

Two remarks are worth emphasizing. First, the proof does *not* require that $\mathbf{x}$ solves a linear assignment objective; it uses only that the chosen allocation maximizes a weighted sum $\sum_i b_i q_i(\mathbf{x})$ over a bid-independent feasible set. Second, the presence of MNL externalities is absorbed into the definition of $q_i(\mathbf{x})$; once a_{ij} are fixed, externalities do not alter the monotonicity argument.

Exploration and randomization. In exploration rounds, the displayed edge (i_t, j_t) is chosen by a rule that is independent of bids, hence each advertiser's click probability $y_i(\mathbf{b})$ is constant in b_i (and therefore monotone). Because the exploration coin (and any randomness in selecting (i_t, j_t)) is drawn after bids are submitted and is independent of bids, we may condition on a realized random seed ω and view the mechanism as deterministic given ω.

Lemma 5.2 (Universal monotonicity under bid-independent randomization)**.** *For each realized random seed ω in round t, the induced deterministic allocation rule is monotone in the sense of Lemma 5.1. Consequently, the unconditional interim click probability $y_{i,t}(\mathbf{b}_t) = \mathbb{E}_\omega[y_{i,t}(\mathbf{b}_t; \omega)]$ is also weakly nondecreasing in $b_{i,t}$.*

Envelope payments and (approximate) DSIC. Given monotonicity, the standard envelope construction yields dominant-strategy incentive compatibility for each deterministic realization of the mechanism. Concretely, for a fixed seed ω and fixed $\mathbf{b}_{-i}$, define $y_i(z) = y_i(z, \mathbf{b}_{-i}; \omega)$. Monotonicity ensures that $y_i(\cdot)$ is almost everywhere integrable and that the payment rule

$$t_i(b_i, \mathbf{b}_{-i}; \omega) \;=\; b_i y_i(b_i) - \int_0^{b_i} y_i(z)\, dz$$

implements truthful reporting as a dominant strategy in the usual single-parameter, quasi-linear sense. Because this holds *for every* ω, the mechanism is in fact *universally* DSIC when the integral is computed exactly.

In the implemented mechanism we approximate the integral using a grid of mesh η, producing a payment $\widehat{t}_i$. This induces an additive incentive loss that we can bound in the worst case by a term linear in η (and, under bid caps, linear in $\bar{v}$ as well).

Proposition 5.3 (ε-DSIC from discretization)**.** *Assume bids are capped in* $[0, \bar{v}]$*. For each round* t *and each realized seed* ω*, the discretized payment rule with step size* η *yields a per-round* $\varepsilon_{IC}(\eta)$*-DSIC guarantee: for every advertiser* i*, every value* $v_i \in [0, \bar{v}]$*, and every deviation* b_i'*,*

$$u_i\big(v_i; v_i, \mathbf{b}_{-i}\big) \;\geq\; u_i\big(v_i; b_i', \mathbf{b}_{-i}\big) \;-\; \varepsilon_{IC}(\eta),$$

where $u_i(v_i; \mathbf{b}) = v_i y_i(\mathbf{b}) - \widehat{t}_i(\mathbf{b})$ *and* $\varepsilon_{IC}(\eta) = O(\bar{v}\,\eta)$*.*

The logic is standard: with exact payments, truthful bidding maximizes utility pointwise in ω; with discretization, the only difference is an additive payment computation error whose magnitude scales with the mesh size. In particular, any potential gain from misreporting must be mediated through this numerical error term, since the envelope identity continues to hold approximately on the discretization grid.

What breaks if estimates depend on bids. The bid-independence of the learning state is not a technical convenience; it is the hinge on which monotonicity (and hence the envelope argument) turns. If the CTR estimator uses data whose distribution depends on current bids, then the mapping $\mathbf{b} \mapsto a_{ij,t}$ becomes endogenous. In that case the objective being maximized in exploitation is no longer of the form

$$\max_{\mathbf{x} \in \mathcal{X}(K)} \sum_i b_i q_i(\mathbf{x}),$$

with $q_i(\mathbf{x})$ fixed; instead one effectively maximizes $\sum_i b_i q_i(\mathbf{x}; \mathbf{b})$, where the quantity function itself depends on bids through the estimator. The key inequalities in the proof of Lemma 5.1 then fail because the comparison between $\mathbf{x}(\mathbf{b})$ and $\mathbf{x}(\mathbf{b}')$ involves two *different* objective functions.

This endogeneity can arise in two practically relevant ways. First, if we update CTR estimates using exploitation impressions/clicks, then which edges are sampled depends on bids; bidders may have incentives to shade bids to change which data are collected (and hence future confidence bounds), generating a dynamic manipulation channel even if each round separately appears "almost truthful." Second, if predicted CTRs incorporate bid-dependent features (or any signals correlated with bids in a way that is strategically controllable), then a bidder can move its own estimated click propensity directly, again invalidating single-parameter monotonicity.

Our mechanism avoids these failures by learning only from single-ad exploration rounds whose selection rule is fixed ex ante and independent of

current bids. This preserves the interpretation of the mechanism as (approximately) a sequence of single-parameter truthful auctions run against a slowly improving, but strategically exogenous, estimate of click primitives.

6 Learning and regret

We now formalize the welfare cost of learning the standalone click primitives and show that the mechanism achieves sublinear welfare regret. The key economic tradeoff is transparent: more exploration accelerates learning (tightening confidence intervals and improving future exploitation), but it also displaces high-welfare allocations in the present because exploration rounds intentionally show at most one ad.

Welfare benchmark and regret. Fix a round t with realized context $c_t = c$. For any feasible matching $\mathbf{x} \in \mathcal{X}(K)$, let $\pi_i(\mathbf{x}; \rho^c)$ denote advertiser i's MNL click probability under the *true* log-odds $\rho^c_{ij} = \log(p^c_{ij}/(1 - p^c_{ij}))$, and define the associated welfare

$$\mathrm{W}(\mathbf{x}; c) \;=\; \sum_{i=1}^{n} v_i \, \pi_i(\mathbf{x}; \rho^c).$$

Let $\mathbf{x}^*_t \in \arg\max_{\mathbf{x} \in \mathcal{X}(K)} \mathrm{W}(\mathbf{x}; c_t)$ be a clairvoyant welfare-optimal matching using the true p^{c_t}. The realized welfare under our mechanism (which uses lower confidence bounds and may explore) is W_t, and the cumulative welfare regret is

$$\mathrm{Reg}(T) \;=\; \sum_{t=1}^{T} \big(\mathrm{W}(\mathbf{x}^*_t; c_t) - \mathrm{W}_t\big).$$

In what follows we take expectations over the i.i.d. contexts, the mechanism's randomization (the exploration coin and any bid-independent exploration design), and the click outcomes.

Concentration from bid-independent exploration. Because exploration rounds display at most one edge (i, j) and the selection of that edge is bid-independent, the platform observes a clean Bernoulli sample with mean p^c_{ij} whenever context c occurs and edge (i, j) is explored. Let $N^c_{ij,t}$ be the number of such samples collected up to time t, and let $\hat{p}^c_{ij,t}$ be the corresponding empirical mean. A standard Hoeffding construction gives, for any confidence level $\alpha \in (0, 1)$,

$$\Pr\left(\left|\hat{p}^c_{ij,t} - p^c_{ij}\right| > \sqrt{\frac{\log(2/\alpha)}{2N^c_{ij,t}}} \right) \le \alpha.$$

Choosing α via a union bound over all (c, i, j) and all $t \le T$ yields a high-probability event on which *all* confidence intervals are simultaneously valid.

Lemma 6.1 (Uniform validity of confidence bounds)**.** *Assume contexts are i.i.d. on a finite $\mathcal{C}$ and exploration feedback for each (c, i, j) is i.i.d. Bernoulli(p^c_{ij}). Define*

$$r^c_{ij,t} \;=\; \sqrt{\frac{\log(4|\mathcal{C}|nmT^2)}{2\max\{1, N^c_{ij,t}\}}}, \qquad \underline{p}^c_{ij,t} = \big[\hat{p}^c_{ij,t} - r^c_{ij,t}\big]_+, \quad \overline{p}^c_{ij,t} = \big[\hat{p}^c_{ij,t} + r^c_{ij,t}\big]_-.$$

Then the event

$$\mathcal{E} \;=\; \Big\{ \forall t \le T,\ \forall c \in \mathcal{C},\ \forall i, j :\ p^c_{ij} \in [\underline{p}^c_{ij,t}, \overline{p}^c_{ij,t}] \Big\}$$

satisfies $\Pr(\mathcal{E}) \ge 1 - O(T^{-1})$ *(in particular,* $\Pr(\mathcal{E}^c) \le T^{-2}$ *after adjusting constants).*

The salient point is not the specific radius but the scaling $r^c_{ij,t} = \tilde{O}(1/\sqrt{N^c_{ij,t}})$, which is what ultimately drives the $\sqrt{T}$-type regret.

A decomposition of welfare regret. We separate regret into three interpretable components.

(i) Estimation conservatism. In exploitation rounds the mechanism solves the winner-determination problem using $\underline{p}^{c_t}_{ij,t}$ (equivalently, $\underline{\rho}^{c_t}_{ij,t}$). Even on the "good" event $\mathcal{E}$, these lower bounds are pessimistic, so the selected matching may differ from the clairvoyant optimum. This is the statistical price of insisting on allocations that are robust to estimation error.

(ii) Exploration cost. In exploration rounds we intentionally forgo multi-slot allocation and show at most one ad, so the welfare in that round can be substantially below $\mathrm{W}(\mathbf{x}^*_t; c_t)$. This term scales with the total number of exploration rounds, i.e. $\sum_{t=1}^{T} \gamma_t$ in expectation.

(iii) Model-bias term. If the CTR estimation system is systematically biased (e.g. due to misspecification or adversarial shifts), then even with abundant data the center of the confidence interval may be displaced from the truth. While our baseline regret bound assumes unbiased Bernoulli samples, we state the degradation from bounded bias explicitly because it will motivate the robustness variants in the next section.

To make (i) quantitative, we use a stability property of MNL choice probabilities: holding the matching $\mathbf{x}$ fixed, the welfare $\mathrm{W}(\mathbf{x}; c)$ is a smooth function of the underlying standalone click parameters, hence small uniform errors in p^c_{ij} translate into small errors in welfare. In particular, on $\mathcal{E}$ the true matrix p^{c_t} lies above $\underline{p}^{c_t}_t$ entrywise, and the per-round welfare loss from using $\underline{p}$ can be bounded by a constant times $\bar{v}$ times an aggregate confidence radius over the at-most-K displayed edges. Summing these radii over time and applying Cauchy–Schwarz yields the familiar $\sqrt{(\#\text{arms})T}$ rate, where the number of arms is $|\mathcal{C}|nm$.

Theorem 6.2 (Welfare regret under exploration-based learning)**.** *Assume (a) contexts c_t are i.i.d. on a finite set $\mathcal{C}$; (b) exploration samples are i.i.d. Bernoulli with means p^c_{ij}; (c) bids are capped in $[0, \bar{v}]$ and (for the welfare benchmark) values satisfy $v_i \in [0, \bar{v}]$. Consider the mechanism that, in exploitation rounds, computes a matching by solving the MNL winner-determination problem on the lower bounds $\underline{p}^{c_t}_{ij,t}$, and, in exploration rounds, displays at most one ad to sample a single (c, i, j) edge. If confidence bounds satisfy $\Pr(\mathcal{E}^c) \le T^{-2}$ as in Lemma 6.1, then the expected cumulative welfare regret satisfies*

$$\mathbb{E}\big[\mathrm{Reg}(T)\big] \;\le\; \tilde{O}\Big(\bar{v}\,\sqrt{|\mathcal{C}|\,nm\,T}\Big) \;+\; O\left(\bar{v}\sum_{t=1}^{T}\gamma_t\right) \;+\; O(\bar{v}),$$

where the $\tilde{O}(\cdot)$ hides polylogarithmic factors in $|\mathcal{C}|, n, m, T$ and the final $O(\bar{v})$ term absorbs the vanishing contribution from the failure event $\mathcal{E}^c$. Moreover, if the CTR estimator is subject to bounded systematic bias in the sense that $\big|\mathbb{E}[\hat{p}^c_{ij,t}] - p^c_{ij}\big| \le \delta$ uniformly, then the above bound degrades by an additive $O(\bar{v}\,\delta T)$ term.

Two comments clarify what drives this bound. First, the $\tilde{O}(\bar{v}\sqrt{|\mathcal{C}|nmT})$ term is the *statistical* component: there are $|\mathcal{C}|nm$ context-position-advertiser primitives to learn, and each exploration round provides only one Bernoulli sample, so the aggregate uncertainty shrinks at the rate $1/\sqrt{\text{samples}}$. Second, the explicit exploration term isolates the *design* choice: the platform can reduce estimation error by exploring more (thereby increasing $N^c_{ij,t}$), but exploration itself is welfare-costly because it uses only one slot.

Choosing γ_t and allocating exploration samples. Theorem 6.2 suggests balancing exploitation quality against the direct cost of exploration. A simple corollary is obtained by choosing an exploration schedule that yields on the order of $\sqrt{|\mathcal{C}|nmT}$ total exploration samples, spread roughly uniformly across (c, i, j).

Corollary 6.3 (A simple exploration schedule)**.** *Suppose exploration in round t occurs with probability $\gamma_t \equiv \gamma$ and, conditional on exploring, the mechanism selects (i, j) uniformly at random among nm edges for the realized context c_t. Taking*

$$\gamma \;\asymp\; \min\left\{1, \sqrt{\frac{|\mathcal{C}|\,nm}{T}}\right\}$$

yields

$$\mathbb{E}\big[\mathrm{Reg}(T)\big] \;=\; \tilde{O}\Big(\bar{v}\,\sqrt{|\mathcal{C}|\,nm\,T}\Big)\,,$$

up to the additional bias term $O(\bar{v}\,\delta T)$ when systematic bias is present.

A second, operationally appealing choice is a decaying exploration schedule (e.g. $\gamma_t \propto t^{-1/2}$), which concentrates exploration early when confidence intervals are widest and gradually transitions to exploitation.

Corollary 6.4 (Decaying exploration)**.** *Under the same conditions as Corollary 6.3, choosing*

$$\gamma_t \;=\; \min\left\{1, \sqrt{\frac{|\mathcal{C}|\, nm}{t}}\right\}$$

and exploring uniformly over edges for the realized context guarantees

$$\mathbb{E}\big[\mathrm{Reg}(T)\big] \;=\; \tilde{O}\Big(\bar{v}\,\sqrt{|\mathcal{C}|\, nm\, T}\Big)\,,$$

again with an additive $O(\bar{v}\,\delta T)$ degradation under bounded systematic bias.

From a policy and engineering perspective, these corollaries highlight a practical takeaway: the platform should scale exploration with the effective dimension $|\mathcal{C}|nm$ of the prediction problem. Richer context taxonomies (larger $|\mathcal{C}|$) and more candidate insertion points (larger m) are beneficial for relevance, but they increase the number of primitives that must be learned and therefore require either more exploration or a longer horizon to achieve the same welfare performance. This observation motivates robustness and misspecification-aware variants, to which we turn next.

6.1 Robustness to manipulation and misspecification

Our learning guarantee in Theorem 6.2 is intentionally stated under a clean statistical model: exploration produces i.i.d. Bernoulli samples whose means are the true standalone click probabilities p^c_{ij}. In practice, the platform typically relies on a larger prediction stack—logging, de-duplication, bot filtering, attribution, and sometimes model-based counterfactual corrections—and each layer can introduce systematic error. Moreover, even when bids do not enter the estimator, the feedback itself may be strategically distorted (e.g. click fraud, coordinated traffic, or template-specific interaction patterns that violate the assumed MNL form). We therefore separate two notions of robustness: robustness to *manipulation of the learning signal* and robustness to *misspecification of the choice model*. The common economic theme is that robustness is not free: it is achieved either by widening the uncertainty set (hence more conservative allocations) or by collecting additional, cleaner data (hence more exploration cost).

A bounded-bias model. We adopt a simple, auditable way to encode systematic prediction error: for each context c and edge (i, j), we allow the estimator to be biased by at most δ in expectation,

$$\big|\mathbb{E}[\hat{p}^c_{ij,t}] - p^c_{ij}\big| \;\le\; \delta,$$

uniformly over time. This captures a variety of operational phenomena: persistent bot traffic that inflates click rates; template effects not captured by the context taxonomy; or systematic underestimation for newly-onboarded advertisers due to cold-start features. While crude, the parameter δ has a concrete interpretation as a *robustness budget* chosen by the platform: larger δ means we are willing to entertain more severe misspecification and therefore act more conservatively.

Under bounded bias, standard concentration bounds around $\hat{p}^c_{ij,t}$ continue to hold for the *mean* $\mathbb{E}[\hat{p}^c_{ij,t}]$, but the mean itself may be displaced from p^c_{ij}. A transparent way to incorporate this into the mechanism is to inflate confidence intervals by δ:

$$\underline{p}^{c,\mathrm{rob}}_{ij,t} \;=\; \Big[\hat{p}^c_{ij,t} - r^c_{ij,t} - \delta\Big]_+, \qquad \overline{p}^{c,\mathrm{rob}}_{ij,t} \;=\; \Big[\hat{p}^c_{ij,t} + r^c_{ij,t} + \delta\Big]_-.$$

This modification leaves the economic structure intact: the bounds remain bid-independent, and solving the winner-determination problem (WDP) on lower bounds preserves bid-monotonicity and hence the envelope-based payment construction. Statistically, however, the presence of δ induces an irreducible per-round welfare gap because even with $N^c_{ij,t} \to \infty$ the true p^c_{ij} may lie δ away from the estimator's center. This is exactly the source of the additive $O(\bar{v}\,\delta T)$ term stated in Theorem 6.2: it is the price of misspecification that no amount of exploration can wash out.

Manipulation of feedback and robust estimators. The bounded-bias assumption can also be viewed as a reduced-form model of manipulation: if an adversary can corrupt a small fraction of observed clicks (or impressions) in a persistent direction, the resulting empirical mean behaves as if it were biased. One operational response is to replace the empirical mean $\hat{p}^c_{ij,t}$ with a robust mean estimator (e.g. median-of-means or trimmed estimators) computed over blocks of exploration samples. Such estimators yield deviation bounds of the same $\tilde{O}(1/\sqrt{N^c_{ij,t}})$ form under heavy tails or ϵ-contamination, at the cost of larger constants and slightly more bookkeeping. In our framework, the critical mechanism-design requirement is simply that the exploration data stream used to form $\underline{p}, \overline{p}$ remain independent of current bids; robustification of the estimator does not alter that independence and therefore does not threaten incentive properties.

A complementary response is *design-based* rather than estimator-based: we can make the exploration policy harder to game by randomizing over templates and positions, throttling repeated traffic patterns, or reserving a small fraction of impressions for "gold" instrumentation with stricter fraud controls. These interventions again fit naturally into our regret decomposition: they effectively reduce δ (less systematic error) but typically increase the opportunity cost of exploration (more constrained or lower-revenue traffic).

Misspecification of the choice model. Bounded bias addresses errors in the primitives p^c_{ij}, but a distinct concern is that the user choice process may deviate from MNL. If the true click probabilities under a slate are not well-approximated by the MNL mapping $\pi(\cdot;\rho)$, then even perfect knowledge of standalone p^c_{ij} does not imply welfare optimality of the MNL WDP. Our stance is pragmatic: we treat MNL as an approximation that delivers tractable optimization and transparent incentives, and we measure performance against the MNL-optimal benchmark. When the platform's policy objective instead requires robustness to model error, we can reinterpret δ more broadly as bounding the discrepancy between the MNL-predicted click probability and the true click probability, uniformly over feasible slates. This yields the same qualitative conclusion: misspecification generates an $O(\bar{v}\,\delta T)$ linear term unless the model class is enriched or the benchmark is weakened.

A robust optimization variant: max–min welfare over a confidence set. Using lower confidence bounds is already a form of robustness: it is pessimistic *entrywise* in the standalone CTR matrix. However, entrywise pessimism does not always coincide with the worst-case welfare under MNL because choice probabilities couple the displayed ads through the denominator. This motivates a more explicit robust optimization variant in which the platform selects a slate to maximize worst-case welfare over an uncertainty set for the primitives.

Fix a round t and context $c = c_t$. Define the (bid-independent) confidence set

$$\mathcal{P}_t(c) \;=\; \prod_{i=1}^{n}\prod_{j=1}^{m}\big[\underline{p}^{c,\mathrm{rob}}_{ij,t}, \overline{p}^{c,\mathrm{rob}}_{ij,t}\big], \qquad \mathcal{R}_t(c) = \{\rho(p) : p \in \mathcal{P}_t(c)\}.$$

A robust welfare criterion chooses

$$\mathbf{x}^{\mathrm{rob}}_t \;\in\; \arg\max_{\mathbf{x}\in\mathcal{X}(K)} \; \min_{\rho\in\mathcal{R}_t(c)} \sum_{i=1}^{n} v_i\,\pi_i(\mathbf{x};\rho).$$

This max–min formulation has two attractive features. First, it makes the role of δ (and statistical uncertainty) explicit: larger uncertainty sets lead to more conservative allocations. Second, because $\mathcal{R}_t(c)$ is constructed from exploration data only, the objective remains monotone in each bid $b_{i,t}$ for a fixed tie-breaking rule whenever the robust WDP is solved exactly.[1]

The limitation is computational. For standard (non-robust) MNL WDP, we can often exploit known reductions and exact solvers; by contrast, the

[1]Intuitively, the inner minimization is bid-independent, so increasing a single bid scales up that advertiser's coefficient in the outer maximization without changing feasible sets. Formally proving monotonicity requires checking that the robust objective is increasing in each weight, which holds for the linear-in-values welfare criterion.

robust counterpart introduces a continuous adversarial choice of ρ coupled across positions through $\pi(\cdot;\rho)$. Even with box uncertainty in p, the worst-case ρ need not occur at an obvious corner of the box for every slate, because lowering one ad's attractiveness can increase another's click share. As a result, solving the robust WDP may require numerical methods.

When numerical methods are needed (and what is practical). There are three regimes worth distinguishing.

(i) Conservative plug-in (no extra numerics). The baseline policy of optimizing on $\underline{p}$ (or $\underline{\rho}$) can be interpreted as a tractable surrogate for the robust max–min problem. It is easy to implement, preserves incentives, and typically produces allocations that are empirically stable.

(ii) Scenario-based robustness (finite reduction). If we approximate $\mathcal{P}_t(c)$ by a finite set of scenarios $\{p^{(s)}\}_{s=1}^{S}$ (e.g. corners, or samples from a posterior), the robust objective becomes

$$\max_{\mathbf{x}\in\mathcal{X}(K)} \min_{s\in[S]} \sum_i v_i\, \pi_i(\mathbf{x};\rho(p^{(s)})),$$

which can be solved by standard mixed-integer formulations with an auxiliary variable representing the minimum. This approach is attractive when S is small (say, dozens), but can become expensive as S grows.

(iii) Continuous robust optimization (inner minimization). If we treat $\mathcal{P}_t(c)$ as continuous, we face a nested optimization problem. A practical approach is to alternate between (a) solving the MNL WDP for a fixed ρ and (b) approximately minimizing welfare over $\rho \in \mathcal{R}_t(c)$ for a fixed slate. The inner problem is smooth in ρ and can be handled by projected gradient methods on the box constraints in p (or equivalently in ρ). This is computationally heavier, and because we approximate the min, we must be careful to preserve bid-monotonicity; in deployments, we would typically fix the numerical tolerance and tie-breaking ex ante and treat residual approximation error as an additional (engineering) source of ε-IC loss.

Overall, we view the robust max–min variant as an optional module: it is most valuable when the platform faces a credible threat of systematic shifts (large δ) and is willing to pay additional computation for stability. In the next section we outline extensions that address richer user models and operational constraints while preserving the same design logic: bid-independent learning, monotone allocation, and welfare guarantees that degrade gracefully with the complexity of the environment.

7 Extensions (brief)

The mechanism we study is intentionally modular: it separates (i) a bid-independent learning pipeline that outputs confidence sets over click primitives from (ii) a monotone allocation rule that optimizes a conservative

welfare objective and (iii) envelope-style payments. This separation makes it relatively easy to extend the framework while keeping the same economic logic. We briefly discuss four directions that matter in assistant monetization systems: alternative user models (cascade), additional platform objectives (trust or policy budgets), nonstationary environments, and an empirical evaluation plan that respects the incentive and learning constraints.

7.1 Cascade-type user models via monotone bucketization

The MNL mapping $\pi(\cdot;\rho)$ is attractive because it yields smooth substitution across ads, but some interfaces behave more like a sequential scan: users examine positions top-down and typically click at most once. A canonical alternative is the cascade model. One convenient parameterization assigns each advertiser–position pair an attractiveness $a^c_{ij} \in (0,1)$ (probability of a click conditional on being examined) and each position a continuation probability $\lambda^c_j \in (0,1)$ (probability the user proceeds after not clicking position j). If allocation $\mathbf{x}$ assigns advertiser $i(j)$ to position j, then the click probability of the ad in position j is

$$\Pr(\text{click at } j \mid \mathbf{x}) \;=\; \Big(\prod_{\ell<j}\big(1-a^c_{i(\ell)\ell}\big)\lambda^c_\ell\Big)\, a^c_{i(j)j},$$

and the expected welfare is the corresponding value-weighted sum. Unlike the MNL objective, this expression is not additively separable across edges because early-position assignments affect downstream examination mass.

From a mechanism-design perspective, the key requirement is that the allocation rule be monotone in each bid holding fixed the confidence objects. A simple way to preserve monotonicity while remaining computationally tractable is to introduce *monotone bucketization*. Concretely, for each context c and position j, discretize estimated attractiveness (or its lower confidence bound) into B ordered buckets,

$$\beta^c_{ij,t} \in \{1,2,\dots,B\}, \qquad \beta^c_{ij,t} \le \beta^c_{i'j,t} \;\Rightarrow\; \underline{a}^c_{ij,t} \le \underline{a}^c_{i'j,t},$$

and then restrict the allocation to satisfy a nested structure (e.g. higher buckets are eligible for higher positions, or a laminar constraint that prevents placing a lower-bucket ad above a higher-bucket ad when both are selected). Under such a restriction, the welfare objective becomes monotone in the bids in the same sense as in the MNL case: increasing b_i increases the coefficient on any feasible assignment involving advertiser i without altering feasibility, so a deterministic exact solver with fixed tie-breaking yields a monotone selection. The bucketization is not merely an engineering hack: it is an economic device that trades off expressiveness for incentive robustness, turning a complex nonseparable objective into a constrained assignment problem whose optimal solution changes in a controlled (monotone) way as weights change.

Statistically, the bucket boundaries can be defined using confidence intervals (e.g. bucketing by $\underline{a}^c_{ij,t}$), ensuring that the discretization remains bid-independent. The cost is approximation error: coarse buckets may sacrifice welfare relative to the fully optimal cascade assignment, and this loss enters regret as an additional modeling term (analogous to the δ-term in Section 6.1). The benefit is that we can keep the same payment construction (up to $\varepsilon_{IC}(\eta)$ from numerical integration) because monotonicity is preserved by design.

7.2 Multi-objective constraints and "trust budgets"

Assistant monetization is rarely a single-objective problem. Beyond welfare (or revenue), platforms impose constraints that proxy for user trust, advertiser quality, latency, or policy compliance. A tractable way to model this is as a knapsack-style budget coupled to the allocation. Let $\kappa^c_{ij} \geq 0$ be a per-impression "trust cost" of showing advertiser i in position j under context c (e.g. measured by predicted user dissatisfaction, policy risk, or a calibrated relevance penalty). In each round, we might require

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \kappa^{c_t}_{ij}\, x_{ij,t} \;\leq\; B,$$

for a fixed budget B, or alternatively enforce a time-average constraint $\sum_{t\leq T}\sum_{ij}\kappa^{c_t}_{ij}x_{ij,t} \leq BT$.

There are two mechanism-compatible ways to handle such constraints. The first is *hard-constraint optimization*: in exploitation rounds we solve the winner-determination problem over the restricted feasibility set

$$\mathcal{X}_B(K) \;=\; \Big\{\mathbf{x}\in\mathcal{X}(K) : \sum_{ij}\kappa^{c_t}_{ij}x_{ij} \leq B\Big\},$$

using conservative CTRs (e.g. $\underline{p}$ or the robust sets from Section 6.1). Since $\mathcal{X}_B(K)$ is bid-independent, exact optimization with fixed tie-breaking continues to be monotone in bids, and the envelope payment remains valid.

The second approach is *soft constraints via bid-independent multipliers*. Introduce a dual variable $\lambda \geq 0$ and solve

$$\max_{\mathbf{x}\in\mathcal{X}(K)} \;\sum_i b_{i,t}\,\pi_i(\mathbf{x};\underline{\rho}_t) \;-\; \lambda_t \sum_{ij}\kappa^{c_t}_{ij}x_{ij},$$

where λ_t is updated online to satisfy the long-run budget (e.g. via projected subgradient ascent on the constraint violation). As long as λ_t depends only on past information and not on current bids, the per-round allocation remains monotone in each $b_{i,t}$. Economically, λ_t acts like an endogenous "shadow price of trust": when the system is overspending the trust budget, the mechanism automatically becomes more conservative in choosing ads, without requiring ad hoc heuristics that may break incentive properties.

7.3 Nonstationarity: sliding-window lower confidence bounds

Real assistant traffic is nonstationary: new advertisers arrive, user intent shifts, and UI templates evolve. A stationary confidence interval that pools all past exploration data can become misleading. A standard remedy is to use sliding-window or discounted estimators. Let W be a window length. For each (c, i, j) we maintain exploration samples only from the most recent W occurrences of context c (or the most recent W time steps, depending on logging granularity), yielding an estimator $\hat{p}_{ij,t}^{c,W}$ and a corresponding radius $r_{ij,t}^{c,W}$. We then define

$$\underline{p}_{ij,t}^{c,W} \;=\; \big[\hat{p}_{ij,t}^{c,W} - r_{ij,t}^{c,W}\big]_+, \qquad \overline{p}_{ij,t}^{c,W} \;=\; \big[\hat{p}_{ij,t}^{c,W} + r_{ij,t}^{c,W}\big]_-,$$

and run the same conservative WDP and envelope payments.

The welfare analysis shifts from static regret to *dynamic regret*, where the benchmark is the sequence of per-round clairvoyant optima under the evolving $p_{ij,t}^{c_t}$. Bounds typically depend on a variation budget such as

$$V_T \;=\; \sum_{t=2}^{T} \max_{c,i,j} \big|p_{ij,t}^{c} - p_{ij,t-1}^{c}\big|,$$

or related drift measures. Intuitively, smaller W tracks changes better but increases statistical noise (larger radii), while larger W reduces noise but lags behind drift. Importantly, the incentive story remains unchanged: the learning rule (windowing, discounting, drift detection) is still bid-independent by assumption, so monotonicity and payments go through. What changes is the platform's chosen exploration schedule γ_t: in nonstationary environments, exploration cannot decay too aggressively, because old data becomes stale. In practice, we would calibrate γ_t (and W) to match observed drift rates, treating the resulting revenue loss as the cost of adaptability.

7.4 Empirical evaluation plan

Finally, we outline an empirical strategy that is aligned with the mechanism's structural constraints. Offline evaluation based purely on logged production traffic is problematic because the allocation is policy-dependent and because bids may change strategically under counterfactual mechanisms. We therefore advocate a hybrid design with three components.

First, implement *instrumented exploration* exactly as assumed by the theory: with probability γ_t, run a bid-monotone exploration rule that shows at most one ad and logs the resulting Bernoulli click sample for a pre-specified (c, i, j) sampling plan. This creates a clean dataset for estimating p_{ij}^{c} (or its cascade analogues) that is insulated from current bids.

Second, evaluate welfare and constraint satisfaction using *online* metrics and audit trails rather than purely offline replay. The relevant quantities—the realized allocations, the computed confidence bounds, the selected slates, and the charged payments—are all observable and can be monitored for monotonicity violations (due to solver tolerances), budget adherence (trust constraints), and stability under drift (windowed intervals). When robust optimization or numerical inner minimization is used, we would pre-commit to tolerances and record approximation certificates, treating any residual as an additional ε-IC term to be bounded operationally.

Third, run controlled experiments that vary (i) the exploration rate schedule γ_t, (ii) the robustness budget δ (or estimator robustification), (iii) the presence of trust constraints, and (iv) the user model (MNL versus cascade bucketization). The theoretical regret decomposition suggests what to measure: exploration opportunity cost $\sum_t \gamma_t$, estimation error through interval widths, and systematic error through observed calibration drift (an empirical proxy for δ). The goal is not merely to maximize short-run revenue, but to validate that the mechanism delivers stable allocations, predictable incentives, and welfare that degrades gracefully as we add the operational constraints that real assistant deployments require.

8 Conclusion: deployable assistant monetization and open problems

Assistant monetization sits at an awkward intersection of mechanism design, online learning, and product policy. The platform must choose what to show (and where), learn user response under rapidly shifting contexts, and do so in a way that advertisers can reason about and that policy teams can audit. The central lesson of our framework is that *deployability* is largely about separation of concerns: we can preserve incentive properties only if the learning pipeline that constructs click primitives is insulated from contemporaneous bids, and only if the allocation rule that consumes those primitives is monotone in each bid with transparent tie-breaking. Once those two ingredients are in place, payments can be attached by an envelope construction (up to a controlled discretization error), and welfare performance can be evaluated through a regret lens that makes the exploration–exploitation tradeoff explicit.

From an economic point of view, the conservative (lower-confidence-bound) optimization step is not merely a statistical device. It is a governance choice: the platform commits to acting as if uncertain edges are *worse* than their point estimates until sufficient bid-independent evidence accumulates. This has two practical implications. First, it reduces the temptation to "chase noise" in sparse contexts, which is a common failure mode in assistant UIs where tail intents and newly introduced templates appear frequently.

Second, it creates a clear operational contract between modeling and mechanism layers: the modeling system can be upgraded (features, architectures, calibration methods) without changing the mechanism, as long as it continues to output confidence objects that are independent of current bids and satisfy stated coverage guarantees. In practice, this kind of contract is what allows auditability: one can test whether exploration sampling plans were followed, whether confidence radii were computed as specified, and whether the allocation solver respected monotonicity and capacity constraints.

The regret guarantees are best interpreted as *comparative statics* rather than literal performance predictions. They highlight which dimensions drive unavoidable learning loss: the effective number of context–advertiser–position primitives that must be estimated, the horizon over which the system is expected to improve, and any systematic misspecification or bias that cannot be eliminated by more data. For product teams, this decomposition is actionable. If the regret bound is dominated by $|\mathcal{C}|$, then UI proliferation and overly fine-grained context definitions are not free: they create statistical fragmentation that must be paid for with exploration. If it is dominated by nm, then the platform should invest in candidate-pruning and representation learning that shares information across advertisers and positions (while preserving bid-independence), because the mechanism can only allocate well among what it can reliably estimate. And if an additive term like δT is empirically large, then the binding constraint is not exploration but calibration and robustness: the platform should treat misspecification and distribution shift as first-order economic risks, not as second-order modeling nuisances.

Several limitations remain salient. Our cleanest incentive statement relies on a deterministic, exactly solved, bid-monotone allocation rule with fixed tie-breaking. Real systems approximate: mixed-integer solvers stop early, neural rankers are nondeterministic, and engineering teams introduce guardrails that can be triggered by bid-dependent signals (e.g., advertiser-level throttles tied to spend). Each such deviation is economically meaningful because it can create non-monotonicities that invalidate envelope payments. A practical deployment posture, therefore, is to treat *approximation error* as a first-class mechanism parameter. One should pre-commit to solver tolerances, record optimality certificates when available, and bound any residual non-monotonicity by an explicit ε term that is monitored continuously. This is not merely paperwork: it is the difference between a mechanism that advertisers can safely treat as price-taking and one that invites adversarial bid experimentation.

Looking forward, two open problems loom particularly large for assistant interfaces.

Multi-click and multi-action user behavior. The MNL and cascade families are natural first approximations because they reduce the outcome

to (at most) a single click event. Assistants, however, increasingly support richer interaction loops: users may click multiple ads, click and then return, or take non-click actions (copy, call, add-to-cart, ask a follow-up) that are economically valuable and correlated with ad load. A multi-click model is not a minor technical tweak; it changes the welfare objective and, crucially, the mapping from allocation to outcomes. If multiple clicks are possible, then the "marginal contribution" of an advertiser can depend on which other advertisers are present in a more intricate way than standard substitution effects. This raises two design questions.

First, can we identify structural conditions under which the resulting welfare objective remains monotone in each bid, so that envelope payments still apply? For example, if the expected total value can be written as a sum of advertiser-specific terms with coefficients that are nonnegative functions of the allocation and independent of bids, then monotonicity may survive even when users can take multiple actions. Second, when such conditions fail, what is the right relaxation? One direction is to design *monotone approximations* of the true objective (e.g., optimizing a conservative surrogate that lower-bounds long-run user value) and treat the gap as a modeling term in regret. Another direction is to move beyond dominant-strategy truthfulness toward weaker equilibrium concepts that may be more realistic in repeated settings, while still providing robust bidding incentives and predictable revenue.

Endogenous generation policies and endogenous candidate sets. Assistant monetization is unusual because the platform is not only choosing an allocation; it is also generating a response. The generated answer, the placement opportunities that appear in it, and even which advertisers are eligible can depend on the conversation state and on upstream retrieval and generation policies. This endogeneity creates a new strategic channel: advertisers may attempt to influence not just *which slot* they win, but *whether a slot exists* or whether their product is retrieved as a candidate. From the mechanism-design perspective, the key challenge is that the mapping from bids to outcomes can become bid-dependent *through the context construction process itself*, violating the separation principle that underpins both monotonicity and clean learning.

A stylized way to see the issue is to imagine that before running an auction the platform chooses a generation action g_t (a template, response length, number of insertion points, or a retrieval depth), which determines the feasible set of positions and candidates. If g_t is optimized using signals that are correlated with bids (directly or indirectly), then increasing a bid can change the feasible set in non-monotone ways, and standard payment formulas no longer apply. Addressing this requires new kinds of commitments. One approach is *two-stage mechanism design*: first choose g_t using a pol-

icy that is explicitly bid-independent (or depends only on past bids through stable aggregates), and only then run a monotone auction over the induced candidate set. Another approach is to model g_t as part of the allocation and impose monotonicity constraints on the joint policy, though this quickly becomes computationally and statistically demanding. Either way, the central economic principle is the same: to claim truthful incentives, the platform must be able to explain, and ideally certify, which parts of the pipeline were insulated from a bidder's report.

These open problems are not academic curiosities; they are exactly where real deployments struggle. As assistants become more proactive and more personalized, the space of contexts grows, drift accelerates, and the line between "ranking" and "generation" blurs. Our contribution is to make one coherent claim in this moving landscape: if we build the system around a bid-independent learning substrate and a monotone allocation core, then we can obtain mechanisms that are simultaneously (approximately) truthful, statistically principled, and operationally auditable. The remaining work is to extend this discipline to richer interaction models and to the generative layer itself, so that assistant monetization can be both economically sound and aligned with the user-trust constraints that ultimately govern long-run value.