Ad Insertion in LLM-Generated Responses

Shengwei Xu[1], Zhaohua Chen[2], Xiaotie Deng[3], Zhiyi Huang[4] and Grant Schoenebeck[1]

[1]*University of Michigan*
[2]*Taobao & Tmall Group of Alibaba, Renmin University of China*
[3]*City University of Hong Kong*
[4]*The University of Hong Kong*

Abstract

Sustainable monetization of large language models (LLMs) remains a critical open challenge. Traditional search advertising, which relies on static keywords, fails to capture the fleeting, context-dependent user intent—the specific information, goods, or services a user seeks—embedded in conversational flows. Beyond the standard goal of *social welfare maximization*, effective LLM advertising requires *contextual coherence* (aligning ads semantically with transient user intent), *computational efficiency* (avoiding user-facing latency), and adherence to *ethical and regulatory standards*, including privacy preservation and explicit ad disclosure. Although recent solutions have explored bidding at the token and query levels, neither category holistically satisfies these constraints.

We propose a framework that resolves these tensions through two decoupling strategies. First, we *decouple ad insertion from response generation* to facilitate pre-screening and explicit disclosure. Second, we *decouple bidding from specific user queries by using "genres"* (high-level semantic clusters) as a proxy. This allows advertisers to bid on stable categories rather than sensitive real-time responses, reducing computational burden and privacy risks. Applying the VCG auction mechanism to this genre-based framework provides approximate guarantees for dominant-strategy incentive compatibility (DSIC), individual rationality (IR), and social welfare. In synthetic experiments with 10^5 advertisers and 100 candidate slots, VCG clears in approximately 1.25 seconds on a consumer-grade laptop. Finally, we introduce an "LLM-as-a-Judge" metric for estimating contextual coherence. Its predictions correlate with mean human ratings at Spearman's $\rho \approx 0.66$ and have a higher correlation with the leave-one-out group mean than 29 of 36 individual raters (80.6%).

Keywords

large language models, ad auctions, mechanism design, contextual coherence

1. Introduction

ChatGPT now serves more than 900 million weekly active users and has over 50 million consumer subscribers [**?**]. Given the considerable cost of training and deploying LLMs—a recent International Data Corporation (IDC) forecast projected $108 billion in U.S. GenAI spending in 2028 [**?**]—sustainably monetizing services on which so many users rely remains challenging.

A natural approach to monetizing LLM services is to integrate ads into LLM responses, similar to current search engines. Indeed, a recent survey [**?**] revealed that 68% of users have used LLMs for "searching to find facts quickly, like a search engine", and 57% of them have used LLMs for "getting information about products and services". For context, Google's advertising revenue exceeded $260 billion in 2024 [**?**]. Thus, introducing ads into LLM responses could help defray the costs of these services and is particularly relevant to conversational commerce.

However, the transition from "Search" to "Chat" is not merely a shift in interface; it presents unique structural challenges. In a search engine, user intent [**?**] is naturally encoded by static keywords, so advertisers can bid on keywords to target a particular intent. These keywords indicate users' state of mind and, in particular, the goods or services they may be seeking. In an LLM chat, the analogous user intent may shift throughout a complex conversational flow. For example, an LLM response about a user's family vacation to New York City might start with a discussion of attractions, then move on to

GenAIECommerce'26: The Third Workshop on Agentic and Generative AI for E-Commerce, co-located with RecSys, September 28, 2026, Minneapolis, MN, USA

† Shengwei Xu and Zhaohua Chen contributed equally.

‡ Shengwei Xu and Grant Schoenebeck are supported by NSF Grant No. 2313137.

accommodations, transportation, and budget considerations. Ideally, an ad about transportation would appear surrounded by text about transportation rather than text regarding accommodations. Placing an ad in the right location not only improves its effectiveness but also makes it less disruptive to the user experience.

This fundamental shift raises three technical challenges:

1. *Contextual coherence*: How do we identify the evolving user intent in the LLM response to provide coherent ads, so that we can align advertisers' welfare with specific users' intents?
2. *Computational efficiency*: How do we insert ads without introducing latency that degrades the real-time user experience?
3. *Social welfare*: How can we design a mechanism that approximately maximizes social welfare, defined as the sum of advertiser and platform utilities, in the absence of explicit, static keywords on which to bid?

In addition, inserting commercial bias into the organic flow is constrained by *practical considerations*. For example, on the *regulation* side, LLM platforms must explicitly mark where ads are inserted, to align with the U.S. Federal Trade Commission (FTC) policy that "advertising and promotional messages that are not identifiable as advertising to consumers are deceptive if they mislead consumers into believing they are independent, impartial, or not from the sponsoring advertiser itself" [**?**]. On the *ethics* side, ads in the LLM response should also be non-deceptive and non-misleading. Further, if advertisers bid at the level of each query or token, they may gain access to potentially sensitive user prompts, creating *privacy* risks.

With these challenges in mind, we aim to answer the following research question:

How can we design a practical framework for integrating ads into LLM responses while preserving advertiser welfare, contextual coherence, and computational efficiency?

1.1. Existing Solutions and Challenges

Traditional search and display advertising models, which rely on fixed ad slots and static user intent, do not generalize well to this domain. LLM responses are dynamic and lack predetermined slots, and an ad's effectiveness depends heavily on its *contextual coherence* with the surrounding text and the user's evolving intent. Conceptually, this resembles *native advertising*, in which ads are woven into a narrative. However, native ads are typically prepared offline and are labor-intensive. Because LLM platforms operate in real time, an automated mechanism must achieve contextual integration with high *computational efficiency*.

In the past two years, there has been a surge of studies on integrating ads into LLM responses. A position paper by **?**] identifies several critical components of LLM advertising, including the question of "what to bid for." To this end, existing works can be roughly classified into two groups, both of which face challenges regarding computational efficiency and practical concerns that our framework addresses.

The first group [**? ? ? ?**] allows advertisers to bid at the *token or segment level*. Here, advertisers bid repeatedly during generation for their content to appear as the next token. This approach introduces significant latency. More critically, as reported by **?**], if LLMs are prompted to blend ads subtly into the organic response, the resulting ads can be difficult to recognize. Because the content is generated in real time by an LLM, it is also susceptible to hallucinations and misinformation.

The second group [**? ? ? ?**] focuses on bidding at the *query or response level*. While it avoids token-level generation issues by inserting predefined ads, bidding on the entire query forces advertisers to value the interaction in aggregate, preventing them from targeting specific moments within a response (e.g., a travel plan containing both "flight" and "hotel" contexts). Furthermore, this approach typically requires advertisers to submit a bid for each specific user query. This also imposes a massive computational burden on advertisers, who must perform real-time semantic analysis to value each unique query effectively.

Crucially, both groups of approaches typically expose sensitive user prompts to advertisers, raising significant privacy concerns.

1.2. Exploratory Survey Results

To explore users' concerns regarding LLM ads, we conducted an exploratory survey with 48 participants in April 2025.[1]

A majority (56.3%) expected ads to appear in LLM responses within a year, whereas only 12.5% thought this would not happen. Despite this expectation, the acceptability of such ads remained low: only 4.2% rated them as acceptable, whereas 64.6% rated them as unacceptable.

To understand this resistance, we analyzed open-ended responses regarding participants' concerns.[2] Three major themes emerged. First, over 37.5% of responses specifically mentioned *trust and objectivity* issues. Users worried about unconscious manipulation through "hidden advertisements":

> *"For some hidden advertisements, due to our trust in AI, we often fail to notice their existence, leading to unconscious biased choices. This is very scary."*

Second, 20.8% of participants expressed concern that using LLMs to generate promotional content may lead to *misinformation or hallucinations* about the advertised product, raising questions about liability and oversight that rarely arise in traditional advertising.

> *"We need to pay more attention to the authenticity of AI-generated ads, and clarify any potential false advertising and who's responsible for it."*

Third, 12.5% emphasized the importance of *contextual relevance* and a seamless user experience.

> *"Will unrelated ads appear in AI response?"*

Together, the survey evidence indicates that users expect LLM ads soon but will only condone them when *clearly disclosed*, *responsible*, and *contextually coherent*. This aligns with legal guidance, reinforces the need to explicitly mark ads in LLM responses, and further motivates our framework.

1.3. Our Contributions

Our primary contribution is a practical framework for integrating ads into LLM responses with high economic welfare, high contextual coherence, and high computational efficiency. En route to this goal, we introduce two layers of decoupling, respectively answering the question of "how ads are displayed" and "what to bid for" raised by **?**].

Concerning how ads are displayed, we **decouple ad insertion from response generation**. In our framework (see **??**), upon receiving a query, an LLM chatbot first generates an organic response to the query, and independently, advertisers compete for the right to insert their pre-submitted ads in different positions in the organic response without further modification (e.g., at the end of a paragraph). Such a decoupling comes with two practical advantages: (1) *Regulation and ethics* concerns related to LLM ads are addressed at their root. Since ad creatives typically stay unchanged for weeks, the service provider can pre-screen and block deceptive content before integrating them into LLM responses. Any ad appearance can also be explicitly marked in the response. (2) The decoupled response-generation and mechanism-design modules can be *independently optimized* with minimal coordination cost. For example, classic auction mechanisms, such as VCG, can be seamlessly migrated into the new scenario.

The other layer of decoupling operates on the bidding interface, where we **decouple bidding from the response context**. Recall that, compared to search ads that rely on keywords, in LLM responses, user intent evolves and thus will be slot-dependent. As we discussed in **??**, it is infeasible for advertisers

[1] Participant demographics are provided in **??**.

[2] Non-English responses were translated into English using Claude 4.5 Sonnet and verified by the authors.

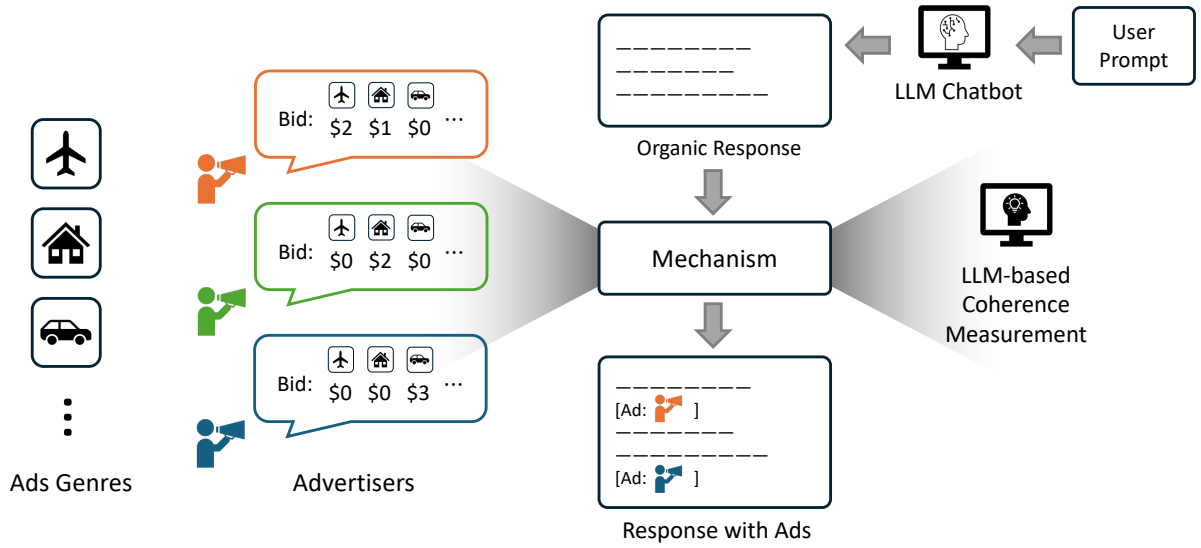

Figure 1: Our framework for ad insertion in LLM-generated responses.

to bid on every possible slot in the response context. Therefore, we introduce the concept of *genres*, high-level semantic clusters (e.g., *hotels*, *airlines*, *food*), as a tractable proxy for the underlying user *intents*. Advertisers bid offline on stable, predefined genres. At inference time, the platform only estimates a *coherence* probability between each genre and each candidate slot in the conversation. Combining (1) an advertiser's genre bids and (2) slot-level genre coherence, we estimate the welfare from assigning that advertiser to a slot by weighting the bids by their coherence probabilities. Conceptually, this is analogous to decomposing a welfare matrix into advertiser–genre and genre–slot factors (see **??**).

Such a design decouples the advertisers from the user's response context, yielding three immediate advantages: (1) *Efficiency:* advertisers no longer need to value every query or token in real time, and the platform runs auctions over a tractable predefined genre set rather than a massive space of queries or tokens; (2) *Privacy:* advertisers bid on coarse categories rather than on potentially sensitive user prompts/responses; (3) Practically, because genres play a role analogous to keywords, the approach can be integrated with existing search-ad infrastructure with limited modification.

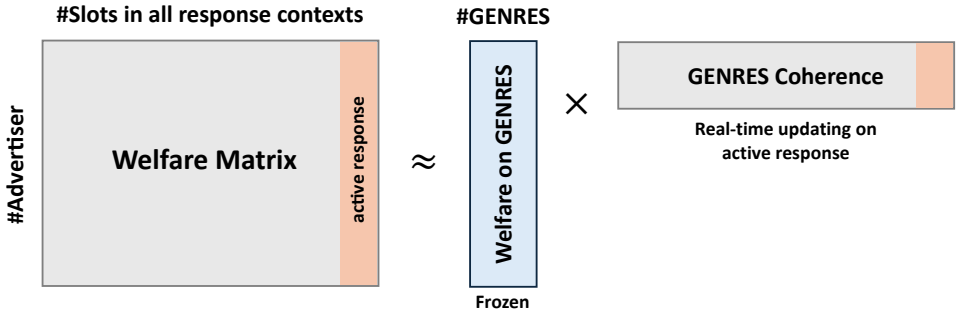

Figure 2: An illustration of how genres work under the matrix decomposition perspective.

However, an estimation error arises because genres may not capture all information about user intent and therefore may not perfectly recover true valuations. This error shrinks as genres become more fine-grained clusters and as coherence estimation becomes more accurate. We now characterize how this approximation affects mechanism design guarantees.

Mechanism design. We analyze the economic properties of this genre-based auction. A core challenge is that genres are a coarse proxy of true user intent; therefore, the platform cannot know the advertiser's true valuation for a specific context perfectly. Despite this information asymmetry, in **??**, we prove that applying the VCG mechanism to the estimated values yields strong theoretical guarantees. Specifically, we show that the mechanism is *approximately* dominant strategy incentive compatible (DSIC) and individually rational (IR). Furthermore, when assuming that advertisers bid truthfully, the mechanism achieves *approximately optimal social welfare*. We derive error bounds showing that the incentive to deviate from truthful bidding is bounded by the granularity of the genres and the accuracy of the coherence estimation. In **??**, we empirically demonstrate that computing the VCG allocation and payment is computationally efficient.

Coherence measurement. For coherence estimation, in **??**, we propose two methods based on recent natural language generation (NLG) evaluation techniques: sentence embeddings [**?** **?**] and LLM-as-a-Judge [**?** **?**].

To evaluate their reliability, in **??**, we conduct empirical studies comparing these coherence signals with human-annotated ratings. Our results show that DeepSeek-R1 and GPT-5-based judges achieve Spearman correlations of approximately 0.66 with the mean ratings of 36 participants. The GPT-5 judge has a higher correlation with the leave-one-out group mean than 29 of 36 individual raters (80.6%). We further observe that the correlation of LLM-as-a-Judge metrics improves consistently with model capability, suggesting continued gains as foundation models advance. We provide all code, experiment materials, and anonymized survey data in the supplementary material.

2. Framework

In this section, we introduce our framework for ad insertion in LLM-generated responses. As shown in **??**, when a user sends a query q to the LLM service, the LLM first generates an ad-free organic response z. Independently, advertisers submit their ad creatives and bidding information to the LLM service provider. Based on the advertisers' values, the mechanism outputs a final response $\tilde{z}$ that preserves the sentences and their order in z while inserting several predefined ad creatives among them.

Crucially, our framework decouples the insertion of ads from the generation of the response. This separation circumvents the risk of hallucinated ad content and ensures transparency, preventing the subtle blending of ads with organic responses.

Our framework consists of three core modules:

1. A valuation model in which advertisers estimate utility using "genres" as a proxy for latent user intents. This design isolates private user queries from advertisers and relieves advertisers of the burden of bidding on every individual query.
2. A contextual-coherence estimator that measures the alignment between a given genre and the organic response.
3. An auction mechanism that combines genre-based bids with the platform's contextual-coherence estimates to determine ad allocations and payments.

Organic response and ad slots. Formally, we model the organic response as a sequence of $M+1$ text segments, $z = z_1 z_2 \cdots z_{M+1}$. For $j \in \{1, \ldots, M\}$, let y_j denote the candidate ad slot between z_j and z_{j+1}. Throughout our framework, we consider a single organic response z to a specific user query. We therefore use j as shorthand for the semantic context at that position in the current response.

2.1. Advertisers' Valuation and Genres

We now model the advertisers' valuation structure. Consider N advertisers, where each advertiser $i \in \{1, \ldots, N\}$ holds an ad creative a_i. Let $w_{i,j}$ be a random variable representing the value derived by the advertiser from an impression at slot j.

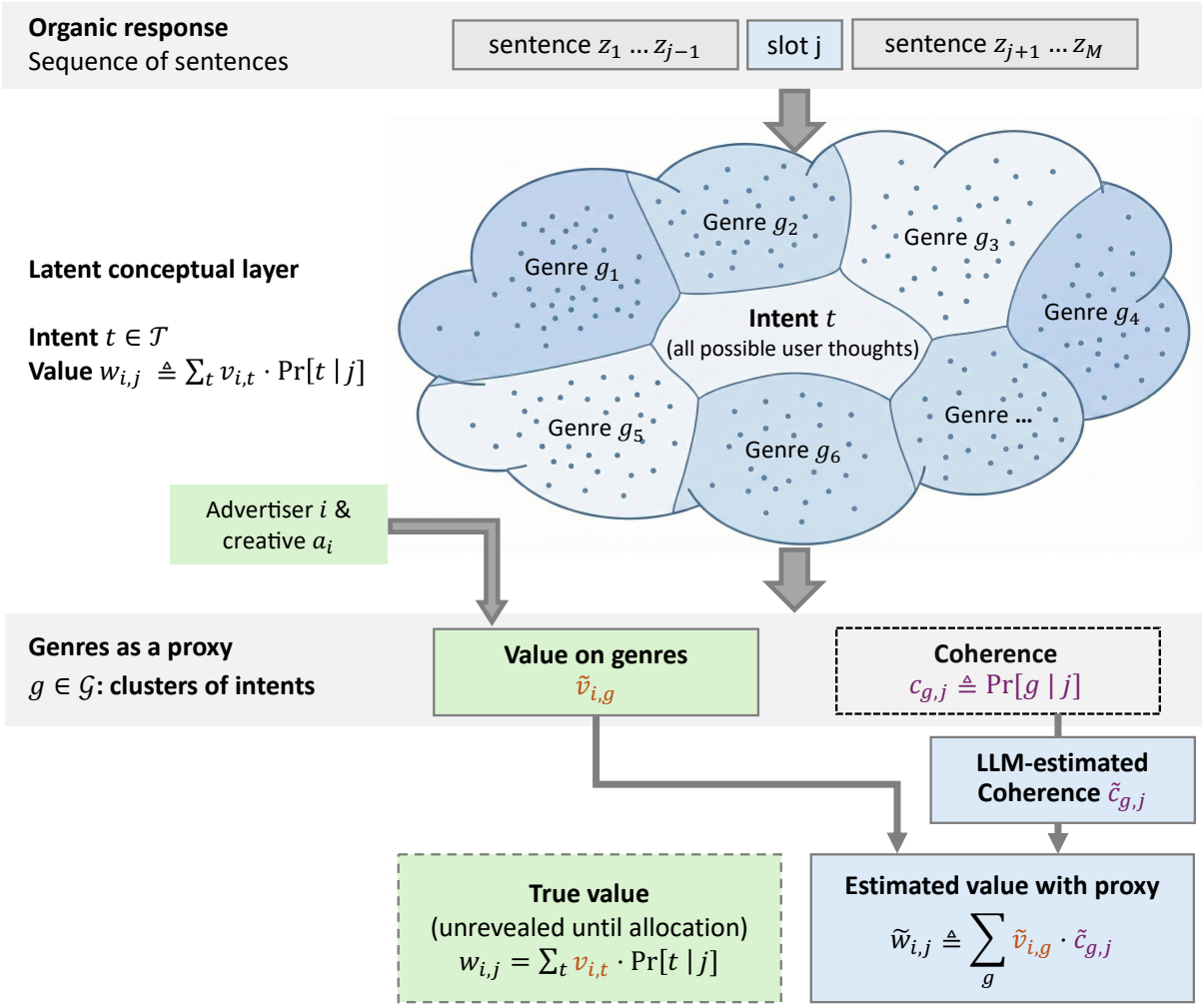

Figure 3: Our model for advertisers' valuation and platform's coherence with genres.
A genre acts as a bucket for multiple intents. The platform uses the average value $\tilde{v}_{i,g}$ as a static proxy, since the true value $v_{i,t}$ is impossible to process due to the extremely large intent space $\mathcal{T}$.
Relying on static proxy averages over multiple intents introduces an estimation error (see **??**). This error decreases as the granularity of the genre set $\mathcal{G}$ increases; in the extreme case where genres equal intents, the proxy matches the true value.

Intents and advertiser valuation. The value $w_{i,j}$ depends on the specific context and the user's intent at the slot. We capture this dependency using the concept of an *intent*, denoted by $t \in \mathcal{T}$, where $\mathcal{T}$ is the set of all possible intents. An intent precisely encapsulates the user's latent thought process while reading the response at a specific slot. However, the textual context at slot j does not fully reveal the user's private latent factors (e.g., budget constraints, current mood). To capture the randomness arising from this lack of full knowledge, we use $\Pr[t \mid j]$ to represent the probability that intent t is active given the context observed at slot j.

Example 1. Consider a user named Alice querying an LLM for a travel itinerary to New York City. Upon reading the transportation section of the response (j), Alice's true intent depends on latent factors unknown to the platform. She might prefer a "first-class flight" if she recently received a bonus or a "Greyhound coach" if she is budget-constrained. The distribution $\Pr[t \mid j]$ captures this uncertainty.

Conditioning on the intent t, the value of an ad creative a_i is independent of the slot index j. Let $v_{i,t} \in [0, \bar{v}]$ represent advertiser i's value given user intent t. Here, $\bar{v}$ is an upper bound on an advertiser's value. Thus, advertiser i's *ground-truth* value at slot j is

$$w_{i,j} \triangleq \sum_t v_{i,t} \cdot \Pr[t \mid j]\,. \tag{1}$$

However, the challenge lies in the cardinality of $\mathcal{T}$, since intents cover the infinite nuances of natural language conversations. Directly estimating $\Pr[t \mid j]$ or bidding on individual $v_{i,t}$ is computationally intractable.

Genres as a proxy. To resolve this intractability, we introduce *genres* as a tractable proxy. We model a genre $g \in \mathcal{G}$ as a set of intents, where $\mathcal{G}$ is a predefined finite set of L genres that forms a partition of the intent space $\mathcal{T}$. Thus, genres do not intersect, and their union covers all intents. This formulation aligns with industry standards for contextual targeting, such as the IAB Content Taxonomy [**?**], which groups similar products into verticals (e.g., food, automotive, and insurance). The probability that genre g aligns with the context at slot j is $\Pr[g \mid j] = \sum_{t \in g} \Pr[t \mid j]$.

We assume that the advertiser knows $\tilde{v}_{i,g} \in [0, \bar{v}]$, the average realized value of its ad creative across a reference distribution of contexts in which genre g has positive probability. In practice, this assumption is supported by the fact that advertisers are required to bid on keywords for search engine ads. Thus, the advertiser's value in **??** can be approximated as

$$w_{i,j} \approx \sum_{g} \tilde{v}_{i,g} \cdot \Pr[g \mid j]\,. \tag{2}$$

Example 2. Consider "Airlines" as a broad genre acting as a proxy for specific user intents, such as "Business Class" and "Economy". While the broad genre remains constant, the underlying intent varies by context: business travel implies a high probability of "Business Class," whereas a budget family vacation implies "Economy." This aggregation causes a valuation error because advertisers value these intents differently. For instance, a low-cost carrier derives high utility from "Economy" but low utility from "Business Class". Since the bid is tied to the coarse genre, it remains static regardless of the slot's specific content. Consequently, the valuation fails to leverage the business or budget signal present in the context, leading to a mismatch between the proxy and true values.

As the granularity of genres increases, the gap between $\sum_t v_{i,t} \cdot \Pr[t \mid j]$ and $\sum_g \tilde{v}_{i,g} \cdot \Pr[g \mid j]$ decreases, since each g captures more information about the context of j. In the extreme case where the number of genres equals the number of intents, each genre represents one intent and the approximation error becomes zero. In **??**, we provide intuition connecting our genre-based model to a low-rank approximation of the welfare matrix. In **??**, we formally discuss how the approximation error impacts the mechanism's truthfulness and social welfare maximization.

2.2. Contextual Coherence

A crucial component of the welfare equation (**??**) is the term $\Pr[g \mid j]$ for all genres g. It captures the probability that genre g aligns with the context of the organic response at position j. Intuitively, if a genre is highly correlated with the text surrounding a slot, this probability increases. We call this probability *coherence* and denote it by $c_{g,j} \triangleq \Pr[g \mid j]$. We further denote $\boldsymbol{c}_j \triangleq (c_{g,j})_{g \in \{1,2,\cdots,L\}}$.

Contextual coherence has been studied in online advertising for decades under names such as *semantic matching* and *relevance modeling* [**? ?**]. Its widespread industrial use suggests that $c_{g,j}$ can be estimated efficiently by building on prior methods.

In our framework, coherence is measured on the platform side since only the platform holds the organic response. We use $\tilde{c}_{g,j}$ to denote the platform's estimation of $c_{g,j}$, and analyze how the estimation error impacts the theoretical bounds in **??**. In **??**, we propose two prototype methods for estimating coherence with modern language models.

2.3. Auction Mechanism

Finally, we formally define the auction mechanism. As illustrated in **??**, the platform conducts the auction using the advertisers' genre-based bids and its own coherence estimates.

Bids and Allocation. Each advertiser i submits a bid vector $\boldsymbol{b}_i \triangleq (b_{i,g})_{g\in\mathcal{G}}^{\top}$, where $b_{i,g} \in [0, \bar{v}]$ represents the advertiser's estimated value for an impression within genre g. We say an advertiser *truthfully* bids if $\boldsymbol{b}_i = \tilde{\boldsymbol{v}}_i \triangleq (\tilde{v}_{i,g})_{g\in\{1,2,\cdots,L\}}^{\top}$. The platform determines an allocation $x_{i,j} \in \{0, 1\}$, indicating whether ad a_i is placed in slot j, and calculates a payment $p_i \geq 0$. Let $K \leq \min\{N, M\}$ be a predetermined integer denoting the number of ads to insert. The allocation must satisfy the following feasibility constraints:

1. Each slot contains at most one ad: $\sum_i x_{i,j} \leq 1, \forall j$.
2. Each advertiser receives at most one slot per response: $\sum_j x_{i,j} \leq 1, \forall i$.
3. Exactly K ads are inserted: $\sum_{i,j} x_{i,j} = K$.

Ground truth utility. Given the allocation and payment, each advertiser i receives a *ground-truth* utility u_i, which is a function of all advertisers' bids $\boldsymbol{b}$ because the allocation $x_{i,j}$ and payment p_i depend on $\boldsymbol{b}$.

$$u_i(\boldsymbol{b}) \triangleq \sum_j x_{i,j} \cdot w_{i,j} - p_i.$$

Proxy utility. Because the true values are unknown to the platform, the advertiser's ground-truth value may differ from the estimated value. With estimated coherence $\tilde{\boldsymbol{c}}_j \triangleq (\tilde{c}_{g,j})_{g\in\{1,\ldots,L\}}^{\top}$ and genre values $\tilde{\boldsymbol{v}}_i \triangleq (\tilde{v}_{i,g})_{g\in\{1,\ldots,L\}}^{\top}$, the *proxy* value that advertiser i obtains in slot j is

$$\tilde{w}_{i,j} \triangleq \sum_g \tilde{v}_{i,g} \cdot \tilde{c}_{g,j} = \tilde{\boldsymbol{v}}_i^{\top} \tilde{\boldsymbol{c}}_j,$$

and the *proxy* utility is therefore

$$\tilde{u}_i(\boldsymbol{b}) \triangleq \sum_j x_{i,j} \tilde{w}_{i,j} - p_i = \sum_j x_{i,j} \tilde{\boldsymbol{v}}_i^{\top} \tilde{\boldsymbol{c}}_j - p_i.$$

2.3.1. Estimation Errors

The gap between the proxy utility $\tilde{u}_i$ and the ground truth utility u_i stems from two sources: the coarseness of the genre proxy and the platform's imperfect coherence measurement. We formally define these errors below; they are used in **??** to analyze the approximation bound for truthfulness and welfare maximization.

Valuation error. Recall that the ground truth value is a sum over intents. We can rewrite this value by grouping intents into their respective genres. Let $v_{i,g,j}$ denote the advertiser's *realized value* for genre g in the specific context of slot j: For $\Pr[g \mid j] > 0$, define

$$v_{i,g,j} \triangleq \mathbb{E}[v_{i,t} \mid t \in g, j] = \sum_{t\in g} v_{i,t} \cdot \frac{\Pr[t \mid j]}{\Pr[g \mid j]}.$$

When $\Pr[g \mid j] = 0$, we set $v_{i,g,j} = \tilde{v}_{i,g}$ by convention; this value has no effect on the ground-truth welfare. This allows us to express the ground truth value in terms of genres:

$$w_{i,j} = \sum_g v_{i,g,j} \cdot \Pr[g \mid j]\,.$$

The *valuation error* arises because the advertiser bids a static value $b_{i,g}$ for the genre (which ideally equals $\tilde{v}_{i,g}$), but the realized value $v_{i,g,j}$ varies by slot (e.g., a "business trip" slot versus a "family vacation" slot within the "Airlines" genre).

Definition 3 (Valuation Error)**.** We define ε_{V} as the maximum normalized deviation between the static genre value and the context-specific realized value:

$$\varepsilon_{\mathrm{V}} \triangleq \frac{\max_{i,g,j} |\tilde{v}_{i,g} - v_{i,g,j}|}{\bar{v}}.$$

Here, $\bar{v}$ is an upper bound on the values $\tilde{v}_{i,g}$.

Coherence error. The second source of error is the platform's ability to estimate the coherence of a genre with the slot context.

Definition 4 (Coherence Error)**.** We define ε_{C} as the maximum $\mathcal{L}_1$ deviation between the estimated coherence vector and the true probability vector over genres:

$$\varepsilon_{\mathrm{C}} \triangleq \max_j \|\tilde{\boldsymbol{c}}_j - \boldsymbol{c}_j\|_1 = \max_j \sum_g |\tilde{c}_{g,j} - c_{g,j}|.$$

3. Incentive Compatibility and Social Welfare Maximization

Our primary goal is to design a mechanism that maximizes social welfare, defined as the sum of the advertisers' true values:

$$W(\boldsymbol{x}) \triangleq \sum_{i,j} x_{i,j} \cdot w_{i,j}.$$

Because payments are transfers from advertisers to the platform, they cancel when advertiser and platform utilities are summed.

However, a fundamental challenge is that the platform cannot observe the true value $w_{i,j}$ derived from latent intents. It can only observe the proxy value $\tilde{w}_{i,j}$ derived from genres. This creates a misalignment between the *proxy* objective optimized by the platform and the *ground truth* value and utility experienced by advertisers (as shown in **??**).

In this section, we analyze how this misalignment affects incentive compatibility and social welfare. The proofs for all results in this section are deferred to **??**.

3.1. Incentive Compatibility

We first examine the incentives for advertisers. Since the platform designs the mechanism based on its own estimates, we define properties regarding the proxy setting versus the ground truth setting.

Definition 5 (DSIC and IR in the proxy setting)**.** We say a mechanism is dominant-strategy incentive-compatible (DSIC) in the proxy setting if truthful bidding maximizes proxy utility, i.e.,

$$\tilde{u}_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) \geq \tilde{u}_i(\boldsymbol{b}_i, \boldsymbol{b}_{-i}), \quad \forall i, \tilde{\boldsymbol{v}}_i, \boldsymbol{b}_i, \boldsymbol{b}_{-i}.$$

We say a mechanism is individually rational (IR) in the proxy setting if truthful bidding yields non-negative utility, i.e.,

$$\tilde{u}_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) \geq 0, \quad \forall i, \tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}.$$

However, advertisers may still deviate if their true utility u_i differs from $\tilde{u}_i$. Due to the platform's imperfect alignment with latent variables, we aim for *approximate* guarantees in the ground truth setting relative to the upper bound on advertiser values $\bar{v} = \max_{i,t} v_{i,t}$.

Definition 6 (Approximate DSIC and IR in the ground truth setting)**.** We say a mechanism is ε-DSIC in the ground-truth setting for $\varepsilon \geq 0$ if truthful bidding maximizes ground-truth utility up to an additive $\varepsilon \cdot \bar{v}$, i.e.,

$$u_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) + \varepsilon \cdot \bar{v} \geq u_i(\boldsymbol{b}_i, \boldsymbol{b}_{-i}), \quad \forall i, \tilde{\boldsymbol{v}}_i, \boldsymbol{b}_i, \boldsymbol{b}_{-i}.$$

We say a mechanism is ε-IR in the ground-truth setting for $\varepsilon \geq 0$ if truthful bidding yields utility of at least $-\varepsilon \cdot \bar{v}$, i.e.,

$$u_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) \geq -\varepsilon \cdot \bar{v}, \quad \forall i, \tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}.$$

The following proposition establishes that if the approximation errors (from values on genres and coherence estimation) are bounded, a DSIC and IR mechanism in the proxy setting remains approximately DSIC and IR in the ground truth setting.

Proposition 1. *If a mechanism is DSIC and IR in the proxy setting, it is (2ε)-DSIC and ε-IR in the ground truth setting, where $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$.*

3.2. Social Welfare Maximization

Because the ground-truth value $w_{i,j}$ is unobservable to the platform, the platform can optimize only the *proxy social welfare*:

$$\tilde{W}(\boldsymbol{x}) \triangleq \sum_{i,j} x_{i,j} \cdot \tilde{w}_{i,j}.$$

We now analyze the optimization of this proxy and derive a bound on the approximation error relative to the ground truth social welfare $W(\boldsymbol{x})$.

As we have shown, for a DSIC and IR mechanism, deviating from truthful bidding can at most yield a utility gain of $2(\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v}$ for each advertiser. Advertisers also do not know the ground-truth quantities $v_{i,g,j}$ and $c_{g,j}$ at bidding time. We therefore evaluate welfare under truthful proxy bidding, the standard benchmark induced by the proxy DSIC mechanism.

3.2.1. VCG Auction

The Vickrey–Clarke–Groves (VCG) auction is known to satisfy DSIC and IR and to maximize social welfare under truthful bidding. In our setting, after receiving all bids $\boldsymbol{b}$ and computing the estimated coherence vector $\tilde{\boldsymbol{c}}_j$ for each j, the platform computes an allocation $\boldsymbol{x}^*$ that maximizes proxy social welfare $\tilde{W}(\boldsymbol{x})$:

$$\boldsymbol{x}^* \triangleq \operatorname{argmax}_{\boldsymbol{x}} \sum_{i,j} x_{i,j} \cdot \boldsymbol{b}_i^\top \tilde{\boldsymbol{c}}_j,$$

$$\text{s.t.} \quad \sum_j x_{i,j} \le 1, \quad \forall i; \qquad \sum_i x_{i,j} \le 1, \quad \forall j;$$

$$\sum_{i,j} x_{i,j} = K.$$

Each allocated advertiser pays the externality that its allocation imposes on the other advertisers:

$$p_i = \max_{\boldsymbol{x}} \sum_{i' \neq i,j} x_{i',j} \cdot \boldsymbol{b}_{i'}^\top \tilde{\boldsymbol{c}}_j - \sum_{i' \neq i,j} x^*_{i',j} \cdot \boldsymbol{b}_{i'}^\top \tilde{\boldsymbol{c}}_j.$$

In particular, when advertiser i does not get allocated, $p_i = 0$.

The main intuition behind the VCG auction is to align each advertiser's objective with the platform's objective—maximizing proxy social welfare—thereby guaranteeing DSIC and IR. To see this, compute each advertiser's proxy utility as

$$\begin{aligned}
\tilde{u}_i(\boldsymbol{b}) &= \sum_j x^*_{i,j} \cdot \tilde{w}_{i,j} - p_i \\
&= \sum_j x^*_{i,j} \cdot \tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{c}}_j + \sum_{i' \neq i,j} x^*_{i',j} \cdot \boldsymbol{b}_{i'}^\top \tilde{\boldsymbol{c}}_j \\
&\quad - \max_{\boldsymbol{x}} \sum_{i' \neq i,j} x_{i',j} \cdot \boldsymbol{b}_{i'}^\top \tilde{\boldsymbol{c}}_j.
\end{aligned}$$

The last term is independent of advertiser i's bid; thus, its objective is to optimize

$$\sum_j x^*_{i,j} \cdot \tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{c}}_j + \sum_{i' \neq i,j} x^*_{i',j} \cdot \boldsymbol{b}_{i'}^\top \tilde{\boldsymbol{c}}_j.$$

Since $\boldsymbol{x}^*$ optimizes $\sum_{i,j} x^*_{i,j} \cdot \boldsymbol{b}_i^\top \tilde{\boldsymbol{c}}_j$, advertiser i's optimal strategy is to bid $\boldsymbol{b}_i = \tilde{\boldsymbol{v}}_i$, and the corresponding utility is non-negative.

3.2.2. Social Welfare Gap

Since the VCG auction that maximizes the proxy social welfare is DSIC and IR in the proxy setting, we can invoke **??** and derive that this mechanism is approximately DSIC and IR in the ground truth setting as well. However, even if we assume advertisers bid truthfully in this case, there remains a gap to the ground-truth social welfare $W(\boldsymbol{x})$. Below, we give a bound on the gap to show that the VCG mechanism behaves well.

Proposition 2. *If the platform uses the VCG auction to maximize proxy social welfare and advertisers bid truthfully, then the gap between the realized social welfare and the optimal ground-truth social welfare* $\max_{\boldsymbol{x}} W(\boldsymbol{x})$ *is at most* $2K \cdot \varepsilon \cdot \bar{v}$*, where* $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$.

3.3. Interpreting the Approximation Guarantees: The Role of Genres

From **????**, we establish that when the platform adopts the VCG mechanism with proxy values, the potential utility gain from strategic deviation is bounded by $2\varepsilon \cdot \bar{v}$ for each advertiser. Furthermore, assuming truthful bidding, the realized ground truth social welfare is suboptimal by at most $2K \cdot \varepsilon \cdot \bar{v}$, with $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$. Consequently, our framework provides conditional approximation guarantees that become informative when both ε_{V} and ε_{C} are small.

While we discuss the estimation of coherence values $c_{g,j}$ and the gap ε_{C} later in **??**, we focus here on drivers of a small valuation gap ε_{V}. Recall that $\varepsilon_{\mathrm{V}} = \max_{i,j,g} |\tilde{v}_{i,g} - v_{i,g,j}| / \bar{v}$. Intuitively, this gap is small if genres provide a good *clustering* of intents, such that the distribution of user intents within a genre remains consistent across slot contexts.

To formalize this, let $D_{g,j}$ denote the conditional distribution of specific intents t given a genre g and slot j:

$$\Pr_{D_{g,j}}[t] \triangleq \frac{\Pr[t \mid j]}{\Pr[g \mid j]}, \quad \forall t \in g.$$

We quantify the variation between these distributions across slots using total variation distance (TVD):

$$d_{\mathrm{TV}}(D, D') \triangleq \frac{1}{2} \sum_{t \in \mathcal{T}} \left| \Pr_D[t] - \Pr_{D'}[t] \right|.$$

We now have the following result.

Proposition 3. *If* $d_{\mathrm{TV}}(D_{g,j}, D_{g,j'}) \le \varepsilon_d$ *for every genre* g *and pair of slots* $j \neq j'$ *satisfying* $\Pr[g \mid j] > 0$ *and* $\Pr[g \mid j'] > 0$*, then* $\varepsilon_{\mathrm{V}} = \max_{i,j,g} \frac{|\tilde{v}_{i,g} - v_{i,g,j}|}{\bar{v}} \le \varepsilon_d$.

4. Coherence Measurement Implementation

In this section, we describe implementations for estimating coherence $c_{g,j} = \Pr[g \mid j]$, the probability that a user's intent belongs to genre g at slot j.

In a mature advertising ecosystem, this probability is typically estimated using extensive historical data (e.g., click-through rates or conversion logs). However, for a novel framework such as ad insertion in LLM responses, such production data may be unavailable, making it potentially intractable to directly measure this probability.

To bridge this gap, our framework introduces a two-step process: First, we compute a raw signal measuring the semantic alignment. We term this estimate the *coherence signal*, denoted by $\tilde{s}_{g,j} \in \mathbb{R}$, which quantifies how well an ad of genre g semantically aligns with the organic context at slot j. Second, we map this signal to a valid probability distribution. This process is often known as *calibration* in machine learning [**?**]. We formally define the calibration step as follows.

Definition 7 (Calibration Function)**.** We define the calibration function $f : \mathbb{R}^L \to \Delta^L$ as a mapping that converts the vector of coherence signals into a probability distribution over genres. For every slot j, the estimated coherence vector is given by:

$$(\tilde{c}_{g,j})_{g\in\mathcal{G}} = f((\tilde{s}_{g,j})_{g\in\mathcal{G}}).$$

In a production environment, f is a learnable module (e.g., isotonic regression [**?**] or temperature-scaled softmax) that fits raw coherence signals to observed user interactions. We focus on deriving high-quality raw signals $\tilde{s}_{g,j}$ and leave data-driven calibration to future deployment.

We now introduce two methods for computing the coherence signal, based on sentence embeddings and LLM-as-a-Judge, respectively. In **??**, we empirically compare these methods' alignment with human-annotated ratings.

Sentence embedding. Sentence embedding models convert a sentence into a numerical vector that represents its semantic meaning. The cosine similarity between two embedding vectors can then be used to measure the semantic similarity between the sentences.

Recall that the organic text is a list of sentences, denoted as $z = z_1 z_2 \dots z_M$. We define the embedding-based coherence signal $\tilde{s}^{(\text{embed})}_{g,j}$ as the sum of the cosine similarities between the ad genre and its two adjacent sentences, minus the similarity between those adjacent sentences. The final term represents the cost of disrupting the original flow:

$$\begin{aligned}\tilde{s}^{(\text{embed})}_{g,j} &= \cos(\text{embed}(z_j), \text{embed}(g)) \\ &\quad + \cos(\text{embed}(g), \text{embed}(z_{j+1})) \\ &\quad - \cos(\text{embed}(z_j), \text{embed}(z_{j+1})).\end{aligned}$$

LLM-as-a-Judge. Recent work has shown that LLMs can act as evaluators of natural-language quality, a paradigm known as "LLM-as-a-Judge." Unlike embeddings, which capture local semantic overlap, this method uses an LLM's reasoning capabilities to evaluate rhetorical flow and contextual relevance.

For each candidate slot j, we present the user prompt, the organic response, the marked slot, and the list of ad genres to the LLM using the structured prompt in **??**. For each genre g, the LLM outputs an estimated coherence signal $\tilde{s}^{(\text{judge})}_{g,j} \in \{1, 2, 3, 4, 5\}$ and a brief rationale.

5. Coherence Measurements' Alignment with Human Evaluation

The two automated coherence measures assess how well ad genres integrate with surrounding organic content. However, a critical question remains:

To what extent do these automated measurements align with human perceptions of ad–context coherence in LLM-generated content?

To answer this question, we compare automated coherence estimates $\tilde{s}_{g,j}$ with human ratings $s_{g,j}$. Our analysis spans a diverse set of organic prompts, ad genres, and insertion positions. To ensure both diversity and representativeness, we select prompts based on prior surveys of LLM user behavior, and choose ad genres grounded in real-world digital advertising expenditure statistics.

Experiment setup. Specifically, we select seven diverse prompts that reflect typical user interactions with LLMs, informed by survey data from over 500 U.S. adults [**?**]: travel planning (reported by 34% of participants), product suggestions (57%), social-gathering planning (23%), creative writing (36%), term explanations (68%), health information (39%), and coding (23%). For each prompt, we generate an organic response using OpenAI's GPT-4o model.

Table 1
Spearman correlations between automated coherence measures and mean human ratings across 310 genre–slot tasks (higher is better).

Base Model	**Qwen3**		
	embedding-0.6B	embedding-4B	embedding-8B
Spearman's ρ	0.2701	0.2838	0.3116

(a) Sentence-embedding-based methods.

Base Model	**Qwen3**		**GPT**			**DeepSeek**	
	8B	32B	4o-mini	4o	5	v3	r1
Spearman's ρ	0.5532	0.5804	0.5856	0.6380	0.6637	0.6636	0.6650

(b) LLM-as-a-Judge methods.

We select 10 high-spend ad genres based on industry spending patterns in digital advertising [**?**]. The selected genres represent sectors with the highest digital advertising expenditure.[3] These genres are presented in alphabetical order in the experiment to mitigate ordering effects.

For each generated text, we identify potential ad insertion points primarily at paragraph boundaries, as these represent typical positions for display advertisements in digital content. For robustness testing, we also select several insertion points within paragraphs that may disrupt the flow of organic content, allowing us to evaluate the metrics across varying degrees of contextual disruption.

Participant recruitment. We recruit 48 participants for our study in China, with 36 completing the entire experiment.[4] According to our pilot study, each session lasts approximately 40 minutes, and participants receive 75 CNY (approximately US$10.41) for their effort.

The demographic data of the participants show that all the participants have completed or are pursuing at least a college degree. For English proficiency, 77.1% of participants reported an advanced level or higher, while the remainder reported an intermediate level. These characteristics suggest that participants could understand the English-language tasks, although they limit the generalizability of our results. The detailed instructions for participants are presented in **??**.

Coherence signal measurement setup. For the sentence-embedding-based measurement, we use Qwen3 embedding models (0.6B, 4B, and 8B), deployed on an NVIDIA A100 80GB GPU.

For LLM-as-a-Judge, we use Qwen3 (8B and 32B), GPT-4o-mini, GPT-4o, GPT-5, DeepSeek-V3, and DeepSeek-R1 through hosted APIs. We include Qwen3-8B to compare embedding-based methods and LLM-as-a-Judge using models of the same size.

In **??**, we provide the prompt used for LLM-as-a-Judge coherence measurement.

Experiment procedure. Across the 7 selected contexts, we have specified 31 possible ad insertion points in total. For each insertion point, participants rate the suitability of inserting each of the 10 ad genres on a scale from 1 to 5, following the instructions in **??**. We define a single evaluation task as a unique (genre, position) pair. This results in 310 tasks in total.

Statistical metric. To represent overall human perception, we compute the average score from all participants for each task. We assess alignment by computing Spearman's rank correlation between the mean human score and each automated coherence measure over all 310 tasks.

[3]To increase category granularity, we divide the retail sector into three sub-genres: packaged food, apparel, and electronics.
[4]Other participants only complete the exploratory survey about online advertising in LLMs, as discussed in **??**.

Result 1: correlation. ?? presents the Spearman correlations between each automated measure and the human annotations. LLM-as-a-Judge correlates more strongly with human judgments than sentence embeddings, including when both use Qwen3-8B. Correlations generally increase from the smaller Qwen3 judges to the proprietary models; DeepSeek-R1 and GPT-5 both achieve approximately 0.66. These results suggest that LLM-as-a-Judge may benefit from future improvements in foundation models.

Result 2: LLM-as-a-Judge aligns more closely with the group mean than most individual raters. To contextualize the correlations, we compare the LLM's performance with that of individual participants. For each participant, we compute the correlation between that participant's ratings and the mean ratings of all other participants. ?? compares the resulting distribution with the GPT-5 judge's correlation. The GPT-5 judge has a higher correlation with the leave-one-out group mean than 29 of 36 individual participants (80.6%). One possible explanation is that modern LLMs are trained to reflect aggregate human preferences, whereas individual judgments contain idiosyncratic variation.

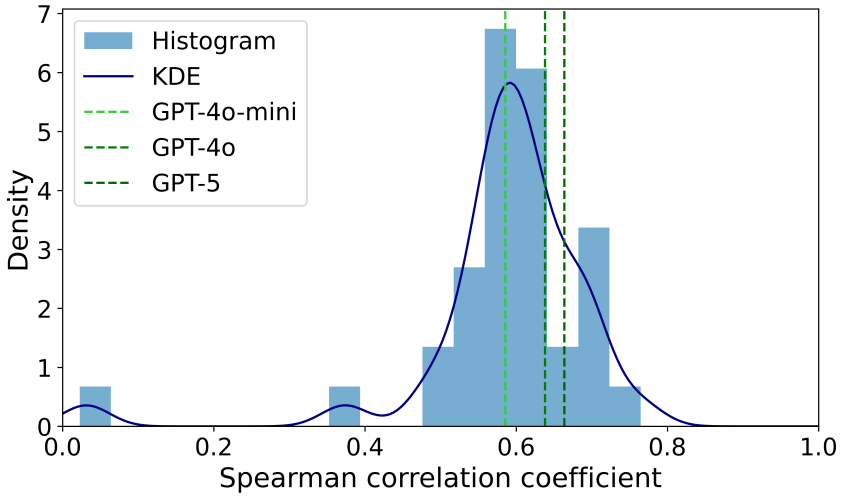

Figure 4: Distribution of leave-one-out individual-to-group Spearman correlations. Dashed lines show three GPT judges; GPT-5 exceeds 29 of 36 individual raters.

Result 3: The GPT-5 judge has a rater equivalence of approximately two. Following ?], we adopt the concept of *rater equivalence*, which quantifies the minimum number of human raters needed to achieve the same expected correlation with the population mean as a given classifier. To estimate this, we compute the average correlation between the mean of Q randomly sampled human participants (denoted as a committee) and the population mean (excluding the sampled committee to avoid bias). We compare these values with the correlation obtained by LLM-as-a-Judge.

In ??, the GPT-5 judge has a rater equivalence of approximately two: its predictions align with the group mean about as well as the mean ratings of two randomly chosen participants. Increasing the committee size substantially increases correlation with the population mean, whereas averaging multiple LLM-as-a-Judge instances yields only modest gains. We hypothesize that this is because different LLMs might be less independent than human raters.

These experiments evaluate the ranking quality of the raw coherence signal, not its calibration as a probability distribution. As discussed in ??, calibrating the signal requires interaction data from a deployed system and remains future work.

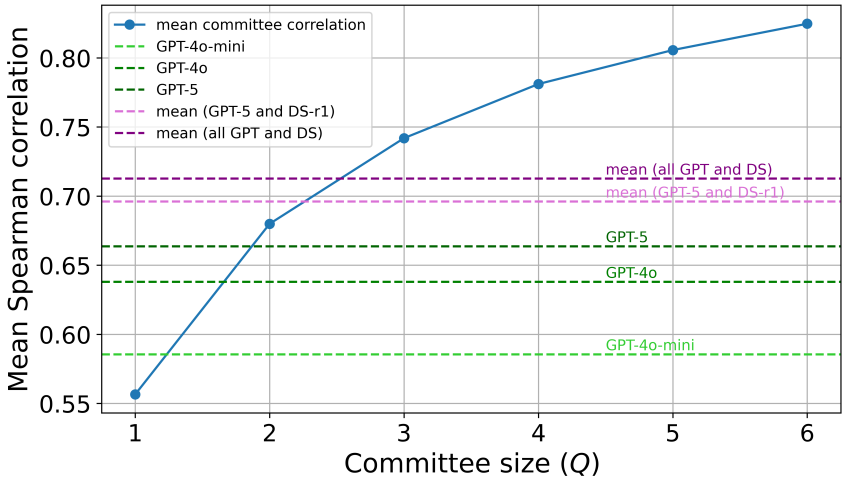

Figure 5: Average committee-to-group Spearman correlation by committee size. GPT-5 aligns with the group mean about as well as a two-person committee.

6. Prototypes

Putting all components together, we implement a prototype of the end-to-end workflow: it accepts genre-based bids and an organic response and produces an ad-augmented response with corresponding payments. Specifically, we use DeepSeek-R1 as the "LLM-as-a-Judge" coherence estimator and min–max normalize its scores to $[0, 1]$.[5] We then apply the VCG mechanism to calculate allocations and payments.

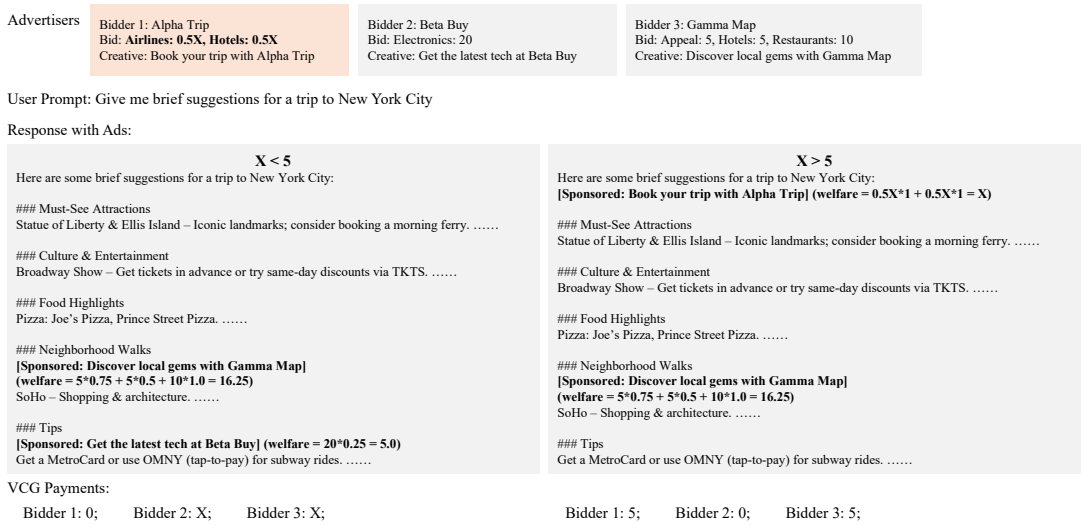

Figure 6: An ad-insertion example using the VCG mechanism ($K = 2$) in a travel-advice scenario. Three advertisers participate: Alpha Trip (Airlines/Hotels), with a uniform bid of $0.5X$ on both genres; Beta Buy, with a bid of 20 on Electronics; and Gamma Map, with bids across Apparel, Hotels, and Restaurants. When $X > 5$, Alpha Trip wins because of high coherence despite its lower raw bids. Payments reflect the welfare of the displaced bidder in each scenario.

?? illustrates a travel planning scenario in which we insert $K = 2$ ads into an LLM-generated organic response. Notably, when Alpha Trip (advertiser 1) increases its bid X above 5, it is allocated an ad slot, despite its bid still being significantly less than Beta Buy's (advertiser 2) bid on Electronics (20). This occurs because Alpha Trip's targeted genres align better with the organic context: Hotels and Airlines both have the maximum coherence score of 1.0 at the beginning of the travel-advice

[5]This heuristic normalization is used only for the illustrative prototype. The approximation guarantees in ?? require calibrated coherence probabilities, which we leave for future deployment.

response. Its resulting proxy welfare is X, which exceeds Beta Buy's welfare of 5. Meanwhile, Gamma Map (advertiser 3) secures the "Neighborhood Walks" slot with the highest proxy welfare (16.25) by aggregating values across multiple high-coherence genres that align with the slot's context.

Finally, the calculated payments follow the VCG rule: winners pay the externality they impose on the system. For example, when Alpha Trip wins ($X > 5$), it displaces Beta Buy (whose welfare is 5) and therefore pays 5.

6.1. Computational Efficiency of VCG

While VCG is often criticized for the computational cost of its maximum-weight matching at scale, the structure of LLM ad insertion mitigates this concern: the number of candidate slots M is naturally bounded by the response length, and the number of inserted ads K is kept small to maintain user engagement. We empirically evaluate this cost in synthetic experiments.

Setup. We use synthetic data to simulate a large-scale ad-insertion scenario. We fix the number of genres at $L = 100$ and the number of inserted ads at $K = 5$. We vary the number of candidate slots $M \in \{20, 50, 100\}$ and advertisers $N \in \{10^3, 10^4, 10^5\}$. For each pair (N, M), we run 100 independent simulation trials.

We solve the matching problem using SciPy's implementation of the Jonker–Volgenant (JV) algorithm [**?**]. All experiments run on a consumer-grade laptop (2020 Apple MacBook Air with an M1 chip).

Synthetic data generation. For advertiser bids, we assume a sparse interest model where each advertiser bids on a limited subset of genres. The size of this subset is determined by a Poisson distribution ($\lambda = 2$), reflecting that most advertisers target only a few specific verticals. For these selected genres, the bid values are drawn uniformly from the interval $[0, 1]$, while valuations for non-selected genres are set to zero.

We also generate a coherence matrix representing the alignment between every candidate slot and ad genre. We draw coherence scores uniformly from $[0, 1]$ and randomly set half of the entries to zero to simulate the common case in which many genres are irrelevant to a slot. We ensure that every slot has non-zero coherence.

Results. **??** presents the mean end-to-end execution time for each configuration, measured from processing the bids and coherence matrix to producing allocations and payments. Runtime scales approximately linearly with the numbers of advertisers N and slots M. In the largest scenario ($N = 10^5$, $M = 100$), VCG clears in approximately 1.25 seconds on the test laptop.

Table 2
Mean VCG execution time in seconds.

	$M = 20$	$M = 50$	$M = 100$
$N = 10^3$	0.0030	0.0064	0.0121
$N = 10^4$	0.0293	0.0659	0.1256
$N = 10^5$	0.3110	0.6946	1.2475

Even in the largest scenario tested ($N = 10^5$ advertisers, $M = 100$ slots, $K = 5$, and $L = 100$ genres), VCG clears in approximately 1.25 seconds on a consumer-grade laptop, with execution time scaling approximately linearly in both N and M. Because LLM inference often takes seconds, these results suggest that the auction itself can be computationally feasible for real-time ad insertion.

7. Related Work

Ad insertion in LLMs. Recent scholarship has focused on integrating ads into LLM responses. A position paper by **?**] outlines critical components in this domain: the modification module (how to display ads), the bidding module (what to bid for), the prediction module (estimating CTR), and the auction module (mechanism design).

On the question of "what to bid for," existing work falls into two broad categories: bidding at the *token or segment level* [**? ? ? ?**] and bidding at the *query or response level* [**? ? ? ?**]. As discussed in **??**, both approaches raise practical concerns, including computational cost, hallucination risk, and privacy leakage. Our framework instead introduces bidding on *genres* to mitigate these tensions.

Other recent mechanism-design papers, including **?**] and **?**], propose mechanisms that may apply to ad auctions in LLM responses but focus primarily on the abstract mechanism-design problem. In contrast, we provide an end-to-end framework that connects bidding, contextual-coherence estimation, allocation, and payment calculation.

Evaluation metrics for natural language generation (NLG). Natural language generation tasks such as machine translation and summarization have traditionally relied on automatic metrics that compare a model's output to human-written reference texts. Classic metrics like BLEU [**?**] and ROUGE [**?**] measure n-gram overlap with reference translations or summaries. In recent years, LLM-based metrics have been proposed to capture deeper semantic quality. For example, BERTScore [**?**] uses contextual embedding similarity. BARTScore [**?**], GPTScore [**?**], and GEM [**? ?**] use pre-trained generative models to score how well a candidate can reproduce the reference or vice versa.

However, inserting advertisements into an LLM-generated response is a new task. We still care about the overall language quality of the augmented response, but the primary evaluation focus is on contextual coherence. Recent work on LLM-as-a-judge evaluation [**? ? ?**] supports using LLMs' reasoning ability to rate outputs according to the given criteria, which is promising for our needs. In addition, we propose a sentence-embedding-based method that uses cosine similarity between ads and their surrounding content as an additional approach for evaluating contextual coherence.

8. Conclusion and Discussion

We present a framework for inserting ads into LLM responses while addressing advertiser welfare, contextual coherence, and computational efficiency. The framework uses two layers of decoupling. First, it separates organic response generation from ad insertion, facilitating ad pre-screening and explicit disclosure. Second, it introduces "genres" that separate bidding from real-time response generation, reducing the privacy and latency costs of sharing full user prompts or responses. Genres thus provide a stable, biddable interface that functions as a counterpart to keywords in search advertising.

Our study has certain limitations that suggest directions for future work.

First, our framework relies on language models to provide accurate coherence measurements. While our LLM-as-a-Judge method aligns with human annotations, further refinements could improve accuracy and efficiency. Future work may explore task-aware embeddings [**?**] tailored specifically to ad–context fit or fine-tune the evaluator via RLHF [**?**] to better capture user intent for ad insertion. Additionally, mapping the raw coherence signals to calibrated probabilities using production data remains a crucial step for real-world deployment. Advances in embedding learning, AI alignment, and generative recommendation systems [**?**] will likely offer useful tools in the future.

Second, using genres mediates the trade-off between efficiency and contextual coherence. Our current model assumes a fixed partition of the user-intent space. Future work could explore adaptive partitioning or hierarchical taxonomies (e.g., *Automotive* $\rightarrow$ *Cars* $\rightarrow$ *Sports cars*) under explicit runtime constraints to determine the genre granularity dynamically, minimizing the valuation error ε_{V} while maintaining scalability.

Third, our human evaluation uses 36 participants from a relatively homogeneous, highly educated population in China and only seven prompt contexts. Broader and more diverse participant populations, prompts, and commercial settings are needed before generalizing the observed correlations.

Finally, ad-supported LLMs may raise a platform-incentive concern: when revenue is tied to impressions or clicks, providers may optimize for engagement (e.g., longer interactions that create more opportunities to show ads) [**?**], potentially distorting the "organic" response. Future work should examine mechanisms that better align platform objectives with user trust and preferences [**?**].

A. Connection to Low-Rank Approximation of the Welfare Matrix

Combining the genre-based valuation proxy from **??** and the contextual coherence, we can interpret our framework as a low-rank approximation of the welfare matrix, as illustrated in **??**.

To formalize this, let $\tilde{\boldsymbol{c}}_j \triangleq (\tilde{c}_{g,j})_{g\in\{1,\dots,L\}}$ denote the estimated coherence vector for slot j, and let $\tilde{\boldsymbol{v}}_i \triangleq (\tilde{v}_{i,g})_{g\in\{1,\dots,L\}}$ denote the genre-based valuation vector for advertiser i. The estimated value for advertiser i at slot j is the inner product $\tilde{w}_{i,j} = \tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{c}}_j$. Let $\boldsymbol{W}_{N\times M} \triangleq (w_{i,j})$ denote the underlying value matrix, whose rows represent advertisers and whose columns represent response contexts (slots). Our framework approximates $\boldsymbol{W}$ with the product of two lower-dimensional matrices:

1. A static valuation matrix $\tilde{\boldsymbol{V}}_{L\times N} \triangleq (\tilde{\boldsymbol{v}}_i)_{i\in\{1,\dots,N\}}$ (genres $\times$ advertisers), which can be elicited by offline bidding and held fixed during LLM inference.
2. A dynamic coherence matrix $\tilde{\boldsymbol{C}}_{L\times M} \triangleq (\tilde{\boldsymbol{c}}_j)_{j\in\{1,\dots,M\}}$ (genres $\times$ current slots), which the platform updates during real-time inference.

The resulting approximation is $\boldsymbol{W} \approx \tilde{\boldsymbol{V}}^\top \tilde{\boldsymbol{C}}$, whose rank is at most the number of genres L. By limiting L to a finite number (e.g., the size of the IAB taxonomy), we reduce the bidding dimensionality from the intractable context space to the tractable genre space.

B. Omitted Proofs

Proposition ??. *If a mechanism is DSIC and IR in the proxy setting, it is (2ε)-DSIC and ε-IR in the ground truth setting, where $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$.*

Proof of **??**. For any bids $\boldsymbol{b}$, we bound the utility gap between two settings as follows:

$$
\begin{aligned}
&|\tilde{u}_i(\boldsymbol{b}) - u_i(\boldsymbol{b})| \\
&= \left| \sum_j x_{i,j} \cdot \left(\tilde{\boldsymbol{v}}_i^\top \tilde{\boldsymbol{c}}_j - \boldsymbol{v}_{i,j}^\top \boldsymbol{c}_j \right) \right| && \text{(where } \boldsymbol{v}_{i,j} \triangleq (v_{i,g,j})_{g\in\{1,2,\cdots,L\}}^\top\text{)} \\
&= \left| \sum_j x_{i,j} \cdot \left[(\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_{i,j})^\top \tilde{\boldsymbol{c}}_j + \boldsymbol{v}_{i,j}^\top (\tilde{\boldsymbol{c}}_j - \boldsymbol{c}_j) \right] \right| \\
&\le \sum_j x_{i,j} \cdot \left[\|\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_{i,j}\|_\infty \cdot \|\tilde{\boldsymbol{c}}_j\|_1 + \|\boldsymbol{v}_{i,j}\|_\infty \|\tilde{\boldsymbol{c}}_j - \boldsymbol{c}_j\|_1 \right] && \text{(Hölder's inequality)} \\
&\le \sum_j x_{i,j} \cdot [(\varepsilon_{\mathrm{V}} \cdot \bar{v}) \cdot 1 + \bar{v} \cdot \varepsilon_{\mathrm{C}}] \\
&\le (\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v}.
\end{aligned}
$$

Therefore, since the mechanism is DSIC in the proxy setting, we have for any $i, \tilde{\boldsymbol{v}}_i, \boldsymbol{b}_i, \boldsymbol{b}_{-i}$,

$$
\begin{aligned}
u_i(\boldsymbol{b}_i, \boldsymbol{b}_{-i}) &\le \tilde{u}_i(\boldsymbol{b}_i, \boldsymbol{b}_{-i}) + (\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v} \\
&\le \tilde{u}_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) + (\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v}
\end{aligned}
$$

$$\leq u_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) + 2(\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v}.$$

This implies that the mechanism is (2ε)-DSIC in the ground truth setting for $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$. Similarly,

$$u_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) \geq \tilde{u}_i(\tilde{\boldsymbol{v}}_i, \boldsymbol{b}_{-i}) - (\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v} \geq -(\varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}) \cdot \bar{v},$$

and the mechanism is ε-IR in the ground truth setting. □

Proposition ??. *If the platform uses the VCG auction to maximize proxy social welfare and advertisers bid truthfully, then the gap between the realized social welfare and the optimal ground-truth social welfare* $\max_{\boldsymbol{x}} W(\boldsymbol{x})$ *is at most* $2K \cdot \varepsilon \cdot \bar{v}$*, where* $\varepsilon = \varepsilon_{\mathrm{V}} + \varepsilon_{\mathrm{C}}$.

Proof of **??**. By the same norm bound used in the proof of **??**, $|\tilde{w}_{i,j} - w_{i,j}| \leq \varepsilon \cdot \bar{v}$ for every advertiser–slot pair. Thus, we have

$$\sum_{i,j} x^*_{i,j} \cdot \tilde{w}_{i,j} - \sum_{i,j} x^*_{i,j} \cdot w_{i,j} \leq \sum_{i,j} x^*_{i,j} \cdot \varepsilon \cdot \bar{v} = K \cdot \varepsilon \cdot \bar{v}.$$

This also implies that $\max_{\boldsymbol{x}} W(\boldsymbol{x}) \leq \max_{\boldsymbol{x}} \tilde{W}(\boldsymbol{x}) + K \cdot \varepsilon \cdot \bar{v}$. Putting them together induces the result. □

Proposition ??. *If* $d_{\mathrm{TV}}(D_{g,j}, D_{g,j'}) \leq \varepsilon_d$ *for every genre* g *and pair of slots* $j \neq j'$ *satisfying* $\Pr[g \mid j] > 0$ *and* $\Pr[g \mid j'] > 0$*, then* $\varepsilon_{\mathrm{V}} = \max_{i,j,g} \frac{|\tilde{v}_{i,g} - v_{i,g,j}|}{\bar{v}} \leq \varepsilon_d$.

Proof of **??**. Recall that

$$v_{i,g,j} = \sum_{t \in g} v_{i,t} \cdot \Pr_{D_{g,j}}[t]\,.$$

Using the property that the difference in expectations is bounded by the TVD scaled by the maximum value, and given $d_{\mathrm{TV}}(D_{g,j}, D_{g,j'}) \leq \varepsilon_d$, we have for any $j \neq j'$:

$$\left|v_{i,g,j} - v_{i,g,j'}\right| \leq \bar{v} \cdot d_{\mathrm{TV}}(D_{g,j}, D_{g,j'}) \leq \varepsilon_d \cdot \bar{v}.$$

Since $\tilde{v}_{i,g}$ is the average value over the reference distribution of contexts in which genre g has positive probability, the distance between any single $v_{i,g,j}$ and the mean $\tilde{v}_{i,g}$ cannot exceed the maximum pairwise distance between $v_{i,g,j}$ and $v_{i,g,j'}$:

$$|\tilde{v}_{i,g} - v_{i,g,j}| \leq \varepsilon_d \cdot \bar{v}, \quad \forall i, g, j.$$

We therefore have $\varepsilon_{\mathrm{V}} \leq \varepsilon_d$. □

C. Experiment Details

C.1. Human Experiment Instructions

?? shows an example of the instructions given to participants in the experiment described in **??**. The full instructions are available in our anonymous repository at https://anonymous.4open.science/r/Ad-Insertion-in-LLM-Generated-Responses-523F/.

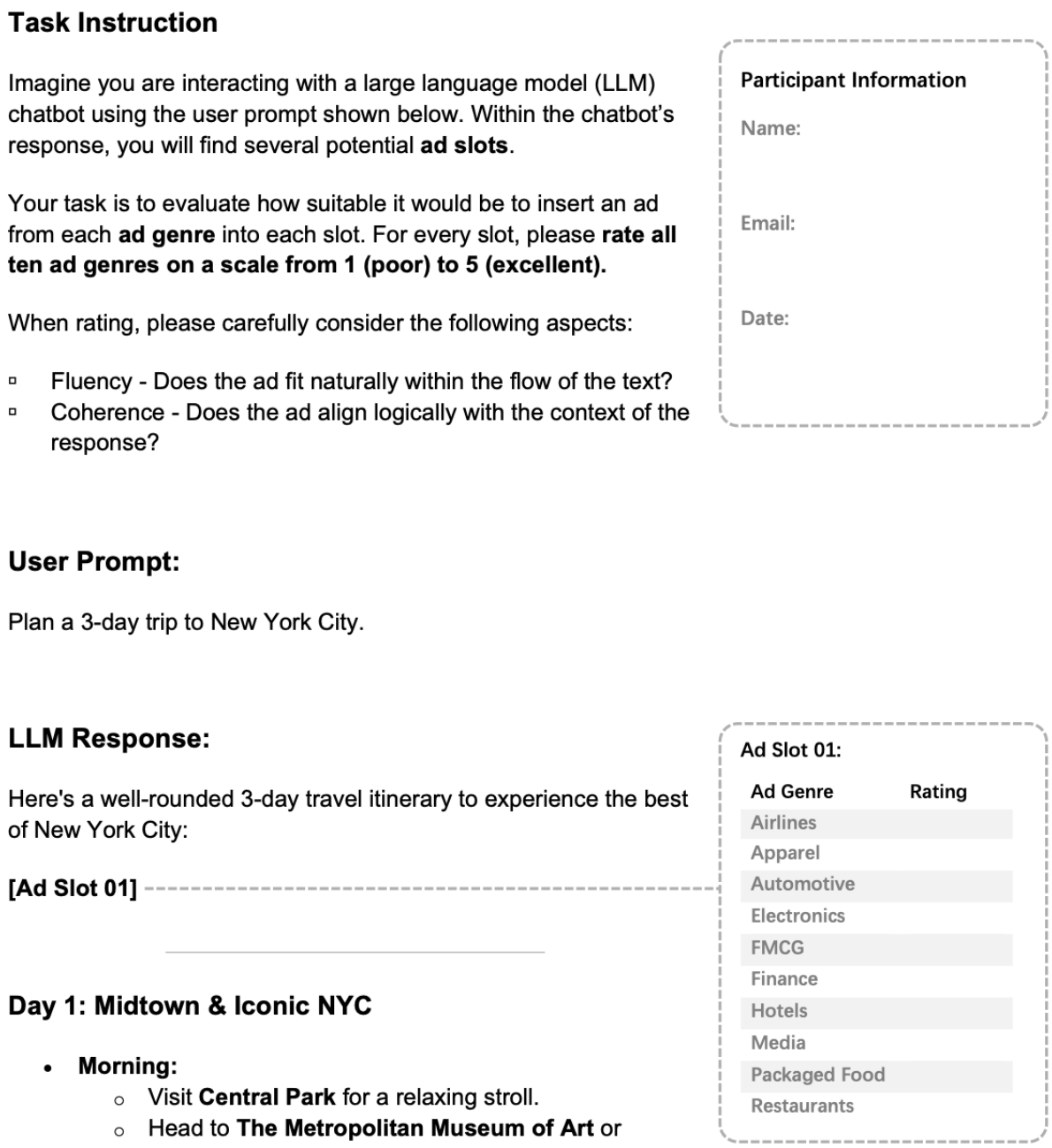

Figure 7: Example instruction for human participants.

C.2. Prompts for LLM-as-a-Judge

This section provides the prompt used for LLM-as-a-Judge coherence measurement in ????.

—— System Prompt

```
You are an expert in digital advertising and user experience. Your task is to rate how suitable different ad genres would be if inserted into a specific location within an LLM-generated response.
```

—— User Prompt

```
Context: {}
User Query: {}

The LLM response contains an ad slot (marked as [Ad Slot]) where an advertisement could be inserted.

```

```
Here is the text surrounding Ad Slot:

{}
[Ad Slot] (THIS IS WHERE THE AD WOULD BE INSERTED)
{}

For each of the following ad genres, rate the suitability of inserting an ad from that genre into this slot on a scale from 1 (poor) to 5 (excellent).

When rating, consider:
- Fluency: Would the ad fit naturally within the flow of the text?
- Coherence: Would the ad align logically with the context of the response?

Here are the ad genres to rate:

Ad Genres:
1. Airlines - Examples: Flight deals, airline promotions
2. Apparel - Examples: Clothing, shoes, accessories
3. Automotive - Examples: Cars, motorcycles, electric vehicles
4. Electronics - Examples: Smartphones, computers, tablets
5. FMCG (Fast-Moving Consumer Goods) - Examples: Personal care products, household items
6. Finance - Examples: Banking services, insurance, credit cards
7. Hotels - Examples: Hotel chains, booking services
8. Media - Examples: Streaming services, social media platforms
9. Packaged Food - Examples: Snacks, beverages, prepared meals
10. Restaurants - Examples: Fast food, cafes, dining establishments

For each genre, provide:
1. A rating from 1-5
2. A brief explanation for your rating

Format your response as JSON with this structure:
{{
  "ratings": {{
    "Airlines": {{ "explanation": "...", "score": X }},
    "Apparel": {{ "explanation": "...", "score": X }},
    ...
  }}
}}

Ensure your ratings are justified based on the context and the natural flow of the text.
```