MAE-AM: Query-driven Multi-Advertisement Embeddings and Auction Mechanism in LLM

Xinpeng Lu[1,2], Heng Song[1,*], Yuanyuan Zhang[1], Xiangyu Shan[1], and Junwu Zhu[1,*]
[1]College of Information Engineering, Yangzhou University, Yangzhou, China
[2]University of Science and Technology of China (USTC), Hefei, China
Emails: {211301216, DX120220097, 231305205}@stu.yzu.edu.cn;
song_heng@foxmail.com; jwzhu@yzu.edu.cn

***Abstract*—Large Language Models (LLMs) contain immense potential for advertising, which prompt a paradigm transition in online advertising from static to generative, embedded display formats campared with traditional search engine. However, the existing advertising systems are not well-suited for meeting the requirements to dynamic reply generation within LLMs. Therefore, the design of effective LLM advertisement auction mechanisms and embedded reply generation becomes significant challenges. In this work, we propose a novel query-driven multi-advertisement embedding and auction system in LLM, which includes two key components: (1) we adopt a Top-q greedy multi-advertisement auction mechanism to determine advertisement embedding lengths, and (2) we incorporate an optimal reply generation scheme based on multi-objective coordination. Specifically, the proposed scheme first integrates auction results with corresponding advertisement content to generate replies via LLM. Subsequently, by adopting multi-objective optimization function for both social welfare and satisfaction, we obtain the best scoring reply that balances the experience between users and advertisers when using LLMs. We theoretically prove that the proposed mechanism satisfies desirable economic properties such as individual rationality and incentive compatibility. Comprehensive empirical results demonstrate that the proposed system outperforms both classical and latest methods in terms of various critical metrics, including social welfare, user and advertiser satisfaction, and overall performance scores. Our code is publicly available at https://github.com/Agentyzu/MAE-AM/.**

***Index Terms*—Computational Advertising; Auction Mechanism Design; Advertisement Generation; Advertisement Pricing.**

I. INTRODUCTION

With the rapid development of digital marketing [1] and Artificial Intelligence Generated Content (AIGC) [2], LLMs have demonstrated exceptional capabilities in information retrieval and personalized content generation [3]–[5], which are widely used in various fields, especially in *online advertising systems*. An online advertising system is a technology platform that uses data and algorithms to accurately deliver advertisements to target users and optimize advertisers' revenue and user experience. Currently, online advertising systems support various display formats, including search engine advertisements, text-based advertisements, and native advertisements [6], [7]. Among these, text-based advertisements are widely adopted due to their advantages of low cost, fast loading and easy customization. Generally, text-based advertisements can be further categorized into enumerative and integrated advertisements, as shown in Fig. 1. Compared with integrated advertisements, *enumerative advertisements* present some advertisement information in a more clear and structured format after a outlined reply, enabling users to quickly locate their concerned content. In contrast, *integrated advertisements* allows the advertisement to be embedded seamlessly into the whole reply, which reduces the sense of dissonance for users' concerned content. Both advertisement formats have their advantages and disadvantages, thereby in actual practice, different user queries should trigger distinct text-based advertisement formats. However, most existing studies focus on a single advertising category and ignore the interaction between query contents and advertisement categories.

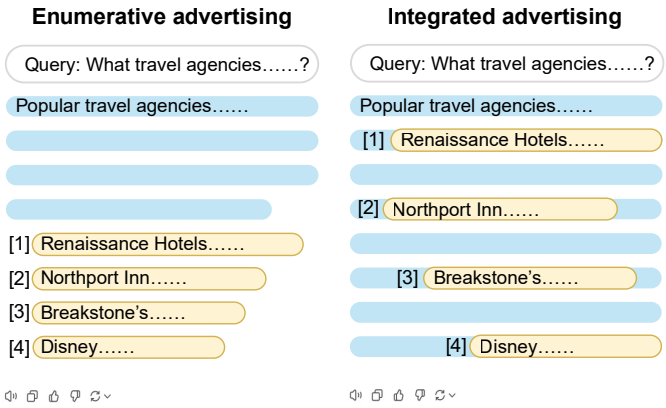


Fig. 1: **Comparison between enumerative and integrated advertisement formats.** For user queries are product or recommendation-oriented (e.g., "Recommend a few reliable travel agencies?"), *enumerative advertisements* provide a clear presentation of advertisement information. In contrast, for conceptual or discussion-oriented queries (e.g., "How does travel help broaden our horizons?"), *integrated advertisements* are more suitable.

Focusing on text-based online advertisement systems, the emergence of LLMs has encouraged advertisers and researchers to explore new paradigms [8], [9]. Unlike traditional search engine advertising, which depends on static keyword matching and fixed display formats [10]–[12], LLMs can embed advertisement content within generated replies seam-

* Corresponding authors: Heng Song and Junwu Zhu.

lessly. This paradigm not only caters to the diverse user needs but also supports complex and contextually rich content. These fundamental differences in output format significantly impact the underlying auction mechanisms. Due to the lack of dynamic content generation paradigms driven by user queries, traditional auction models designed for static search engine advertisements are not directly applicable to LLM-based advertisement auctions [13], [14]. This discrepancy also gives rise to two key research topics: (1) designing LLM-based advertisements auction mechanisms, and (2) effectively embedding advertisements LLM-generated replies [15], [16]. Specifically, the primary objectives of this problem are: *first*, to optimize the allocation of words for embedding advertisements in replies based on advertisers' bidding, allocation and payment rules [8], [17], [18]; and *second*, to incorporate predefined advertisement content into replies that ensures effective placement while maintaining a high-quality user experience.

Recently, significant progress has been made in aforementioned topics such as advertisement content generation [19], keyword extraction [20], and sponsored question-answering [21]. Nevertheless, LLM-based online advertising systems still encounter several challenges: *from a theoretical perspective*, embedded advertisements often vary in format and length, whereas traditional auction mechanisms are primarily designed for static content with predefined formats, thereby limiting their adaptability to variable content. *From the practical perspective*, the performance of LLM-based systems depends not only on the quality of the generated content but also on factors such as the number of advertisement slots and content length. Insufficient design of these factors may reduce the allocation efficiency and reply relevance, thereby impairing system performance.

In this paper, we propose MAE-AM system, a query-driven **M**ulti-**A**dvertisement **E**mbedding and **A**uction **M**echanism in LLM, which employs a multi-advertisement auction mechanism to efficiently allocate the length of words and seamlessly embeds advertisement content into generated replies. The main contributions of this work are summarized as follows:

- We propose MAE-AM system, a query-driven multi-advertisement embedding and auction mechanism in LLM. The system adopts a Top-q greedy auction mechanism to allocate slots and dynamic words based on their corresponding prominence. By designing specific allocation and payment rules to maximize social welfare. Additionally, we prove that the mechanism satisfies critical economic properties, including individual rationality and incentive compatibility. This provides a theoretical foundation to design high-performance systems.
- To embed advertisement content into LLM-generated replies, we propose an optimal reply generation scheme based on multi-objective coordination. The scheme first decouples the mechanism from the generation process by integrating the auction results and content into prompts that guide the LLM's reply. To deliver the best replies, we next evaluate the candidate replies within the LLM community for multi-objective, including social welfare, users and advertisers satisfaction.
- To evaluate the effectiveness of our proposed system, we conduct comprehensive experiments on two datasets: ADGEN and ATVI. The results show that MAE-AM consistently outperforms existing methods in terms of social welfare, user and advertiser satisfaction, and overall score. Moreover, we provide an in-depth analysis of key factors, such as the number of slots and candidate advertiser, content length. Especially, we analyze the performance impact of different advertisement formats and query types, further offering actionable insights for optimizing system deployment.

The rest of this paper is organized as follows: Section II introduces related work; Section III presents the system model and problem formalization; Section IV provides methodology in detail and proves the economic properties; Section V presents extensive experiments to verify the effectiveness of the system. Finally, Section VI concludes the paper.

II. Related Works

With the growing potential of LLMs in the field of mechanism design and advertisement, their applications have drew significant attention and achieved notable progress. In this section, we briefly review the related work, focusing on LLMs and mechanism design, computational advertising.

A. LLMs and Mechanism Design

Recently, researchers have explored the intersection of game theory, mechanism design, and LLMs, highlighting their significant potential in multi-agent interactions, resource allocation, and personalized services [22]–[26]. For instance, literature [27] conceptualizes LLMs as agents in two-player games, exploring the relationship between strategy selection during training and game-theoretical methods, thereby highlighting the role of game theory in understanding the learning dynamics of LLMs. Similarly, literature [28] introduces a game-theoretic framework for multi-agent planning with LLMs, where models dynamically adjust objective function parameters to optimize the efficiency and effectiveness of collaborative decision-making among agents. Dutting et al. [29] proposed a token auction model that couples auction mechanisms with LLM-generated content, effectively aggregating preferences of multi-LLM. It provides theoretical support for token auctions. Additionally, to optimize the fine-tuning and adaptation processes of LLMs, literature [30] proposed a multi-parameter mechanism design framework. This framework employs customized training and payment rules to incentivize participants to report truthful preferences, significantly improving both fine-tuning performance and model efficiency.

B. Computational Advertising

Driven by machine learning and other technologies, computational advertising has continuously optimized advertising strategies through the analysis and modeling of massive data, which has significantly improved the revenue of online advertising systems [31], [32]. In the search engine advertising model, advertisers usually set keywords based on their

products or services, provide content for display and set their own bid. The effect of this model is highly dependent on the experience and investment of advertisers, and it is difficult to make full use of the personalized needs and dynamic characteristics of users. However, online advertising systems can more accurately match user needs and provide higher quality advertising content by leveraging LLMs' strong capabilities in integrating content. So as to gradually change the traditional advertising practice mode [33]–[35].

For example, LLMs enhance auction processes by capturing users' semantic preferences and contextual information, enabling more accurate valuation models. Consequently, literature [36] introduced the SPVA framework, which incorporates LLMs for personalized valuation, facilitating automated advertiser valuation and bidding. This approach offers more flexible and efficient solutions for advertising systems. Soheil etc. [15] explore the potential of LLM in online advertising systems and proposes an LLMA framework with four modules; literature [21] designed sponsored question answering platform based on generalized second price auction. In this platform, LLM is used to combine individual advertisement with user answers to enhance the relevance of content to user needs. Similarly, Dubey [8] extends traditional position auctions and optimizes summaries generation by introducing a modular framework for large models and click-through rate prediction; literature [9] proposes a paragraph auction scheme based on Retrieval Augmented Generation (RAG) technique. This scheme takes advantage of the context generation of LLM, which can not only generate high-quality advertisement content, but also optimize it according to user preferences.

III. Preliminaries

In this section, we first propose our design of MAE-AM (see Fig. 2) and then show some fundamental definitions to ensure the effectiveness of auction mechanism. TABLE I summarizes the major notations used in this paper.

TABLE I: Notations and their definitions.

Notations	Definitions
i, j, k	Index for advertiser, LLM, and slot.
n, m, q	Number of advertisers, LLMs, and slots.
$\boldsymbol{p^{norm}}, \boldsymbol{p^{ctr}}$	Normalized and base click-through rate.
$v_{i,j}, b_{i,j}$	Valuation and bid of ad_i on the LLM_j.
$x(\boldsymbol{b}, \boldsymbol{p^{ctr}})$	Allocation rule of the mechanism $\mathcal{M}$.
$p(\boldsymbol{b}, \boldsymbol{p^{ctr}})$	Payment rule of the mechanism $\mathcal{M}$.
X_i^a	Original advertisement contents of ad_i.
X_j^u, X_j^{LLM}	Base and candidate replies generated by LLM_j.
$Prom_k$	Prominence allocated to slot k.
$L_{\max}$	Maximum word length limit in the system.
$\sigma(k)$	Index of the advertiser allocated to slot k.
SW	Social warfare generated by the mechanism.
S^a, S^u	Satisfaction of advertisers and users.

A. System Model

1) MAE-AM *Framework:* The MAE-AM includes four main roles: 1) *User*: Individual who send query to the system; 2) *Advertisers*: A group of merchants who display their product advertisements through the system; 3) *LLM Community*: A group of LLMs (e.g. the ChatGPT series) to generate replies; 4) *Meta-LLM*: The core model responsible for *managing the operation and decision-making* of the system. The flow of the MAE-AM system consists of the following six main steps:

- **Step 1:** The user sends a query to MAE-AM, which may contain complex semantics, making it difficult for advertisers to submit bids solely based on keyword matching.
- **Step 2:** The Meta-LLM generates a *base reply* to the user's query, which only includes the base content of the query without any embedding.
- **Step 3:** Advertisers calculate bids based on the similarity between their *original advertisement content* and the base reply, indicating their demand for the placement.
- **Step 4:** Advertisers submit their bids, original content, and base click-through rate to the Meta-LLM.
- **Step 5:** Meta-LLM outputs the payments and the *final reply* through allocation and payment mechanism (§IV-A) and reply generation scheme (§IV-B), respectively.
- **Step 6:** Final reply and payments are fed back to the user and advertisers respectively, with the final reply including both the base reply and embedded advertisements.

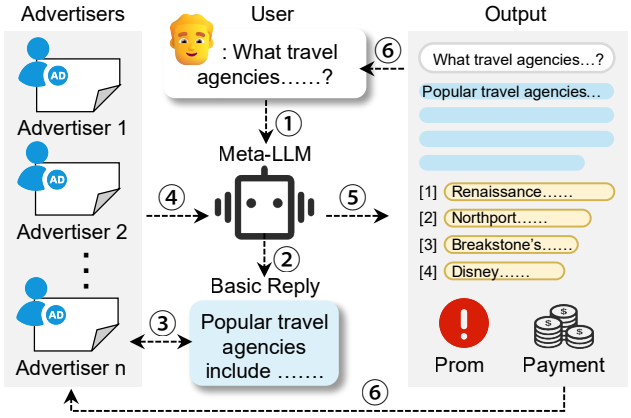


Fig. 2: Schematic diagram of the MAE-AM system flow.

2) Problem formalization: In this section, we formally define the multi-advertisement auction model through LLM community. Let n advertisers $ad_i, i \in [n]$ participate in a multi-advertisement auction across m large language models $LLM_j, j \in [m]$, each placing their advertisement on the models. Each LLM's reply contains q advertisement slots $k \in [q], q < n$[1], where each slot has a normalized Click-Through Rate (CTR) $\boldsymbol{p^{norm}} \in [0,1]^q$, reflecting the impact of its order on the click-through rate (e.g., advertisement positioned later usually have a lower click-through rate). Let the original advertisement content from the ad_i be $\{X_i^a\}_{i\in[n]}$, with a base CTR $\boldsymbol{p^{ctr}} = (p_i^{ctr})_{i\in[n]}$. The base CTR represents

[1]Due to the limited space for content in LLM reply, the number of slots q is pre-set by the system based on available display word limit and is typically smaller than the number of advertisers n.

the inherent attractiveness of the original content, which is typically predicted based on historical data.

To obtain embedding opportunities, advertisers need to bid to the LLM. Let the valuation and bid of advertiser ad_i on LLM_j be $v_{i,j}$ and $b_{i,j}$, respectively. Thus, the valuation matrix is $V = (v_{i,j})_{i\in[n],j\in[m]}$, and the bid matrix is $B = (b_{i,j})_{i\in[n],j\in[m]}$. Note that we assume there is no competition among advertisers within the LLM community, and each column vector $\boldsymbol{b_j} \in B$ represents the bids of all advertisers for a single LLM_j. In this setting, the bid matrix B can be viewed as the combination of bid vectors in multiple *single-parameter environments* [37], i.e., $B = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m]$. The allocation of advertisers in the LLM community can thus be formed by combining the allocation results in each LLM. To simplify the notation, we focus on the auction for a single LLM and omit the second dimension that indicates the LLM's index in the bid notation. Specifically, we use $\boldsymbol{b} = (b_i)_{i\in[n]}$ instead of $b_{i,j}$.

In terms of auctions, after receiving the advertisers' bids $\boldsymbol{b}$ and their base CTR $\boldsymbol{p^{ctr}}$, the auction mechanism $\mathcal{M} = (x, p)$ determines both allocation and payment. This mechanism consists of two fundamental components: the allocation rule x and the payment rule p, which are defined as follows:

- The allocation rule $x : \boldsymbol{b} \times \boldsymbol{p^{ctr}} \to [0, 1]^q$ determines the embedding positions and length of the content.
- The payment rule $p : \boldsymbol{b} \times \boldsymbol{p^{ctr}} \to \mathbb{R}^q_{\geq 0}$ determines the price advertisers pay to the system.

Hence, the utility $u_i(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ of advertiser ad_i is the difference between their valuation and payment, i.e.,

$$u_i(\boldsymbol{b}, \boldsymbol{p^{ctr}}) = v_i x_i(\boldsymbol{b}, \boldsymbol{p^{ctr}}) - p_i(\boldsymbol{b}, \boldsymbol{p^{ctr}}). \tag{1}$$

In terms of reply generation, when a user submits a query Q to the system, each LLM in the community first generates base reply $\{X^u_j\}_{j\in[m]}$ based on the query, i.e., $X^u_j \sim Gen_j(Q)$. Subsequently, using the base reply X^u_j and the original advertisement content $\{X^a_i\}_{i\in[n]}$, the LLM further generates candidate replies X^{LLM}_j, i.e.,

$$X^{LLM}_j \sim Gen_j(X^u_j, \{X^a_i\}_{i\in[n]}).^2 \tag{2}$$

Here, the function $Gen(\cdot)$ represents the content generation process, which generates a corresponding reply based on the query and generation formats. Finally, the system selects the optimal reply from these candidate replies.

B. Fundamental Definitions

To ensure the effectiveness, the auction mechanism $\mathcal{M}$ should aim maximize metrics such as revenue or social welfare while satisfying key economic properties defined as follows:

*Definition 3.1 (**Individual Rationality, IR**):* For an advertiser ad_i, $i \in [n]$, and given the bids $\boldsymbol{b_{-i}}$ of the other advertisers, if the utility of advertiser ad_i is non-negative when bidding truthfully at its valuation v_i, i.e.,

$$u_i((v_i, \boldsymbol{b_{-i}}), \boldsymbol{p^{ctr}}) \geq 0, \tag{3}$$

then the auction mechanism $\mathcal{M} = (x, p)$ satisfies IR.

*Definition 3.2 (**Dominant Strategy Incentive Compatibility, DSIC**):* For an advertiser ad_i, $i \in [n]$, if truthful bidding always yields the highest utility for advertiser ad_i, regardless of the bids of other advertisers $\boldsymbol{b_{-i}}$, i.e.,

$$u_i((v_i, \boldsymbol{b_{-i}}), \boldsymbol{p^{ctr}}) \geq u_i((b_i, \boldsymbol{b_{-i}}), \boldsymbol{p^{ctr}}), \tag{4}$$

then the auction mechanism $\mathcal{M} = (x, p)$ satisfies DSIC.

The individual rationality property ensures that advertisers will participate in the auction only if they can achieve non-negative utility. And the dominant strategy incentive compatibility property guarantees that truthfully reporting their valuations is the optimal strategy for advertisers[3].

*Definition 3.3 (**Monotonicity**):* An allocation rule $x(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ is monotonic if, for two advertisers ad_i and $ad_{i'}$ ($i \neq i'$), their bids b_i and $b_{i'}$ satisfy $b_i > b_{i'}$, then their allocations satisfy:

$$x((b_i, \boldsymbol{b_{-i}}), \boldsymbol{p^{ctr}}) \geq x((b'_i, \boldsymbol{b_{-i}}), \boldsymbol{p^{ctr}}), \tag{5}$$

where $\boldsymbol{b_{-i}}$ denotes the bids from all advertisers except ad_i.

The monotonicity of the allocation rule ensures that increasing a bid does not lead to a reduced allocation, thereby encouraging advertisers to bid truthfully based on their actual valuations. This is essential for designing DSIC mechanisms.

In the reply to the embedded advertisements, the actual click-through rate of them are influenced by both position and length. Due to limited display words, it is impractical to embed all advertisers uniformly, resulting in competition among advertisers for embedding opportunities within a reply. To address the challenge of reasonably allocating the lengths of embedding, we introduce the concept of *prominence*, analogous to the allocation probability in item auctions, which is formally defined as follows [8]:

*Definition 3.4 (**Prominence**):* The allocation rule $x_i(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ is explicitly represented by the prominence of advertisers, denoted as $\boldsymbol{Prom} \in [0, 1]^q$, where $\sum_{k=1}^{q} Prom_k = 1$.

The definition of prominence is interpreted as the *proportion of embedded advertisements*, with their lengths determined based on this proportion. Specifically, the length of an embedded advertisement for ad_i is given by $L_{\max} \cdot Prom_i$, where $L_{\max}$ represents the maximum words length limit supported by the system beyond the base reply. Notably, since the LLM-generated replies are assumed to align with the results of the auction mechanism, the lengths of embedded advertisements are also consistent with the monotonicity of the allocation rule.

Since prominence is defined over continuous allocations, the allocation result for each advertiser is no longer binary (i.e. fully embedded or not embedded at all). Instead, it allows for more flexible adjustments in the extent of advertisement embedding. In this sense, the prominence can be regarded as a **continuous relaxation** of traditional advertisement ranking.

[2]Here, the superscripts u, a, and LLM of X represent the base reply, the original advertisement content, and the LLM-generated reply, respectively.

[3]In subsequent discussions, since the auction mechanism satisfies DSIC, we do not distinguish between valuation v and bid b in notation.

IV. METHODOLOGY

In this section, we provide a detailed explanation of MAE-AM's fifth step (see Fig. 3), which comprises *two primary stages* of Meta-LLM: (i) Allocation and Payment Mechanism, and (ii) Reply Generation Scheme.

A. Allocation and Payment Mechanism

This section introduces the first phase of the Meta-LLM, where propose a Top-q greedy multi-advertisement auction mechanism. Advertisers begin by submitting bids to the system. Existing approches, which ignore query context [8] or rely solely on query text [9], often perform poorly in practice (§V-B) due to the ambiguous and context-dependent user queries. For example, the term "apple" in a query might refer to a phone brand or a fruit, depending on the context. Therefore, an important design is to evaluate the similarity between the content and the base reply to determine the bid. This design improves the relevance, thereby enhancing both the effectiveness of the placements and user experience.

Specifically, the bid b_i of advertiser ad_i is calculated based on the semantic similarity between the base reply X_j^u and the original ad content X_i^a. By encoding the text using a pre-trained language model (e.g. BERT [38]) and computing the similarity $sim(X_j^u, X_i^a)$ between the feature vectors, the bid is normalized to lie between 0 and 1.

Next, we will provide a detailed explanation of the auction mechanism $\mathcal{M}$. The goal of advertisers is to maximize social welfare SW, which is given by:

$$\begin{aligned} \max SW &= \sum_{k=1}^{q} b_{\sigma(k)} \cdot p_{\sigma(k)}^{ctr} \cdot p_k^{norm} \cdot f(Prom_k) \\ \text{s.t.} \quad & \sum_{k=1}^{q} Prom_k = 1, \quad Prom_k > 0, \ \forall k \in [p] \end{aligned} \tag{6}$$

where $\sigma(k) = i$ denotes the embedding of advertiser ad_i's content in the slot k, the function f is a fixed function of the prominence, and it is positive in its domain to measure the effectiveness of the placement (e.g., advertisement exposure).

To solve this optimization problem, we can break it into two subproblems: how to optimize the embedding positions of advertisements in the reply and how to allocate the embedding length reasonably, i.e., calculating the optimal values of $Prom$. Therefore, we propose a Top-q greedy multi-advertisement auction mechanism (see Algorithm 1).

The proposed mechanism first sorts advertisers in descending order based on both their bid and base CTR (lines 1–3), and the Top-q advertisers with higher bids are prioritized for selection, where the advertiser embedded in slot k is denoted by index $\sigma(k)$. Subsequently, the product of $b_{\sigma(k)} \cdot p_{\sigma(k)}^{\text{ctr}}$ and the normalized CTR p_k^{norm} is computed to obtain the effective cost per mille ($ecpm$) for slot k. Based on this, leveraging the allocation rule properties under the social welfare maximization (§IV-C), the prominence of each advertisement is determined (lines 4–6). Here, the function g is defined as the inverse of f', and the parameter ν is calculated as detailed in the proof of Theorem 2.

Algorithm 1: Top-q Greedy Multi-Advertisement Auction Mechanism

Input: Advertisers' bids $\boldsymbol{b}$, base click-through rates $\boldsymbol{p^{ctr}}$, normalized click-through rates $\boldsymbol{p^{norm}}$, and the number of slots q

Output: Advertiser $ad_{\sigma(k)}$ embedded in slot k, prominence $Prom_k$, and payment $p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$, $k \in [q]$

1 **Initialization:** Calculate $b_i \cdot p_i^{ctr}$ for each advertiser $ad_i, i \in [n]$
2 Sort $b_i \cdot p_i^{ctr}$ in descending order, with the sorting slot $\sigma(k) \leftarrow \arg\operatorname{sort}(b_i \cdot p_i^{ctr})$
`// Allocation Rule:`
3 **for** $k = 1$ to q **do**
4 $\quad ecpm_k \leftarrow b_{\sigma(k)} \cdot p_{\sigma(k)}^{ctr} \cdot p_k^{norm}$
5 $\quad Prom_k \leftarrow g\left(\nu / ecpm_k\right)$
6 **end**
`// Payment Rule:`
7 **for** $k = 1$ to q **do**
$\quad$ `// If f is logarithmic function`
8 $\quad w_{-k} \leftarrow \sum_{k'=1}^{q} ecpm_{k'} - ecpm_k, t \leftarrow ecpm_k / w_{-k}$
9 $\quad p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) \leftarrow w_{-k}[\ln(1+t) - \frac{t}{1+t}] / (p_{\sigma(k)}^{ctr} p_k^{norm})$
10 **end**
11 **return** $\sigma(k), Prom_k, p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}), k \in [q]$

Finally, the mechanism computes the payment according to the payment rule (lines 8–10), where w_{-k} represents the sum of $ecpms$ for all slots except slot k, and t denotes the ratio of the $ecpm$ for slot k to w_{-k}. The payment $p_{\sigma(k)}$ at slot k is calculated as described in line 9. The derivation and proof of this formula are provided in Theorem 3.

B. Reply Generation Scheme

This section presents the second stage of the Meta-LLM, in which an optimal reply generation scheme based on multi-objective coordination is proposed.

As shown in Algorithm 2, the process begins by ensuring that the LLM adheres to allocation results and content. This is achieved through stage-wise optimization by *decoupling* the auction mechanism from embedded advertisement generation, with prompts serving as the dynamic *interface* between the two parts. The intuition behind this design is that advertisers cannot directly control the content and format generated by the LLM. As a result, all strategy formulation is completed before the reply generation. Specifically, the auction results are merged with the original content to create a *new prompt*, which subsequently guides the LLM community to generate X_j^{LLM} (lines 1-6). Here, $tpl(\cdot)$ represents the prompt construction function that synthesizes the auction results and content into a specified format.

Next, after the LLM community generates candidate replies, we calculate the bilateral satisfaction of both advertisers and users (lines 7-13). User satisfaction S^u is measured by calculating the semantic similarity $sim(X_j^{LLM}, X_j^u)$

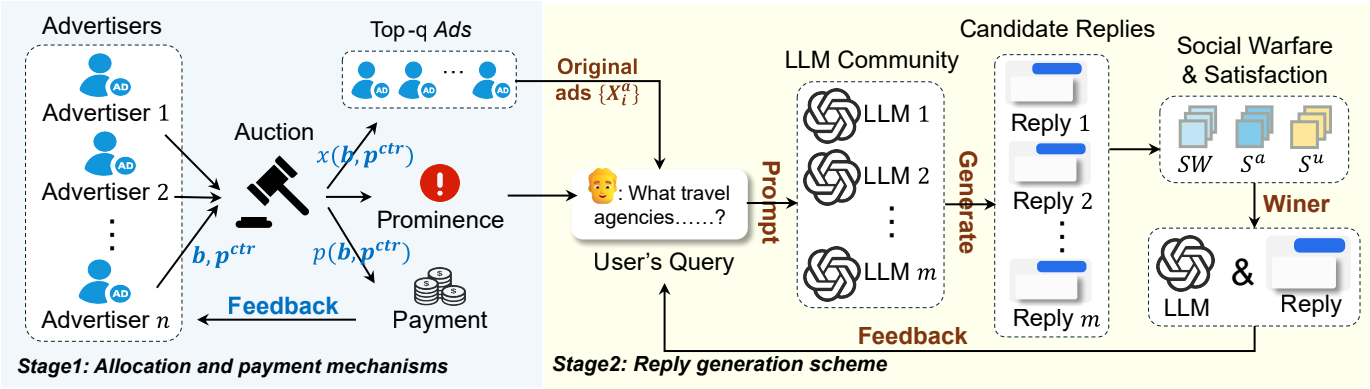


Fig. 3: Schematic diagram of the *two primary stages* of Meta-LLM in MAE-AM.

between the candidate reply X_j^{LLM} and the base reply X_j^u. Similarly, advertiser satisfaction S_k^a is measured by $sim(X_j^{LLM}(\sigma(k)), X_{\sigma(k)}^a)$, where $X_j^{LLM}(\sigma(k))$ represents the content of slot k in the candidate reply. If the content is not embedded, the similarity is 0. The overall advertiser satisfaction S^a is the average satisfaction of all advertisers whose contents are embedded, i.e., $S^a = \sum_{k=1}^{q} S_k^a/q$.

Algorithm 2: Optimal Reply Generation Scheme Based on Multi-Dimensional Objective Coordination

Input: Advertisers $ad_{\sigma(k)}$ embedded in slot k; advertiser prominence $\boldsymbol{Prom}$; original advertisement content $\{X_i^a\}_{i\in[n]}$; base replies $\{X_j^u\}_{j\in[m]}$; and the number of slots q

Output: Final reply X, winning LLM_{j^*}

1 **Initialization:** Parameters θ_1, θ_2 and $Prompt$
// Construct a new prompt:
2 **for** $k = 1$ to q **do**
3 $\quad Prompt \leftarrow Prompt + tpl(X_{\sigma(k)}^a, Prom_k)$
4 **end**
// Calculate metrics and score:
5 **for** $j = 1$ to m **do**
6 $\quad X_j^{LLM} \leftarrow Gen_j(Prompt, X_j^u, \{X_{\sigma(k)}^a\}_{k\in[q]})$
7 $\quad S_{sum}^a \leftarrow 0$
8 $\quad S^u \leftarrow sim(X_j^{LLM}, X_j^u)$
9 $\quad$ **for** $k = 1$ to q **do**
10 $\quad\quad S_{sum}^a \leftarrow S_{sum}^a + sim(X_j^{LLM}(\sigma(k)), X_{\sigma(k)}^a)$
11 $\quad$ **end**
12 $\quad S^a \leftarrow S_{sum}^a/q$
13 $\quad Score(X_j^{LLM}) \leftarrow \theta_1 S^u + (1-\theta_1)S^a + \theta_2 SW_j$
14 **end**
// Output the final reply:
15 $j^* \leftarrow \arg\max_j Score(X_j^{LLM})$
16 $X \leftarrow X_{j^*}^{LLM}$
17 **return** LLM_{j^*}, X

Finally, the scoring function integrates multiple objectives, including user satisfaction, advertiser satisfaction and social welfare, to calculate the comprehensive score for each candidate reply (lines 14–15):

$$Score(X_j^{LLM}) = \theta_1 S^u + (1-\theta_1)S^a + \theta_2 SW_j, \quad (7)$$

where θ_1 and θ_2 are coordination factors used to adjust the weights of the multiple objectives. When prioritizing user experience, θ_1 should be set to a higher value, whereas in scenarios focusing on advertisers, θ_1 can be appropriately reduced. The winning model LLM_{j^*} is determined by $j^* = \arg\max_{j\in[m]} Score(X_j^{LLM})$, and the candidate reply $X_{j^*}^{LLM}$ with the highest score will be selected as the final reply and presented to the user.

C. Theorem Proving

In this section, we provide a detailed analysis and formally prove the properties implied by the proposed auction mechanism $\mathcal{M}$ based on the aforementioned algorithm.

Theorem 1: Under the condition of maximizing social welfare, if the prominence function f is a **concave function**, then the allocation rule $x(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ inherently satisfies *Prominence Allocation Monotonicity (PAM)*.

Proof: Under Algorithm 1, if for any two slots $k, k' \in [q]$, the following condition holds: $ecpm_k > ecpm_{k'} \Leftrightarrow Prom_k > Prom_{k'}$, then satisfy the *Prominence Allocation Monotonicity (PAM)*. It is straightforward to demonstrate that the allocation mechanism implemented by Algorithm 1 satisfies monotonicity (see Definition 3.4) with respect to bids.

Next, we reformulated the problem of social welfare maximization as: $SW = \sum_{k=1}^{q} ecpm_k \cdot f(Prom_k)$. To solve this optimization problem, we construct the Lagrange function by introducing the Lagrange multiplier ν, as follows:

$$\mathcal{L} = \sum_{k=1}^{q} ecpm_i f(Prom_i) - \nu(\sum_{k=1}^{q} Prom_k - 1). \quad (8)$$

Assuming f is continuously differentiable, we take the partial derivative of $\mathcal{L}$ with respect to $Prom_k$:

$$\frac{\partial \mathcal{L}}{\partial Prom_k'} = ecpm_i f'(Prom_k) - \nu. \quad (9)$$

Let g denote the inverse function of f', such that $g(f'(x)) = x$. Setting derivative to zero yields: $Prom_k = g(\frac{\nu}{ecpm_k})$. The value of ν is determined by the constraint

$$\sum_{k=1}^{q} g\left(\frac{\nu}{ecpm_k}\right) = 1. \tag{10}$$

Since ν is the same for all k, we further derive:

$$ecpm_k f'(Prom_k) = ecpm_{k'} f'(Prom_{k'}), \forall k, k' \in [q]. \tag{11}$$

If the mechanism satisfies PAM, then for $ecpm_k > ecpm_{k'}$, it follows that $Prom_k > Prom_{k'}$ and $f'(Prom_k) < f'(Prom_{k'})$. Since $f'(Prom)$ is a monotonically decreasing function, it follows that f is a concave function. ■

If the function f is concave, the marginal effect of ad embedding diminishes as prominence increases. This property discourages the excessive concentration of resources on a small number of advertisers, promoting a more balanced allocation. The result aligns with intuitive reasoning — for example, increasing the length of a text from 10 to 20 words has a greater impact than extending it from 100 to 110 words.

Theorem 2: Under the maximization of logarithmic social welfare[4], $x(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ ensures *proportional allocation*.

Proof: From Theorem 1, we know that when $f(Prom_k) = \ln(Prom_k) + C$, then $g(Prom_k) = f'(Prom_k) = \frac{1}{Prom_k}$. In this case, $\nu = \sum_{k=1}^{q} ecpm_k$, and

$$Prom_k = g\left(\frac{\nu}{ecpm_k}\right) = \frac{ecpm_k}{\sum_{k'=1}^{q} ecpm_{k'}}. \tag{12}$$

Thus, the maximum logarithmic social welfare is given by:

$$SW_{\max} = \sum_{k=1}^{q} ecpm_k \left(\ln\left(\frac{ecpm_k}{\sum_{k'=1}^{q} ecpm_{k'}}\right) + C\right). \tag{13}$$

Since the allocation of advertiser $ad_{\sigma(k)}$ is proportional to $ecpm$, the allocation rule satisfies proportional allocation. ■

Theorem 3: Under the maximization of logarithmic social welfare, for any advertisement slot $k \in [q]$, the allocation rule $x_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ and payment rule $p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ for advertiser $ad_{\sigma(k)}$ are defined as follows:

$$x_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) = \frac{ecpm_k}{\sum_{k'=1}^{q} ecpm_{k'}}, \tag{14}$$

$$p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) = \frac{w_{-k}}{p^{ctr}_{\sigma(k)} p^{norm}_k}\left[\ln(1+t) - \frac{t}{1+t}\right],$$

where

$$w_{-k} = \sum_{k'=1}^{q} ecpm_{k'} - ecpm_k, t = \frac{ecpm_k}{w_{-k}}. \tag{15}$$

Under these conditions, the auction mechanism $\mathcal{M}$ satisfies both incentive compatibility and individual rationality.

Proof: According to Myerson's lemma [37], the auction mechanism $\mathcal{M}$ is *incentive compatible* if the payment rule $p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ can be derived from the monotonic allocation rule $x_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ using the following formula:

$$p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) = \int_0^{b_{\sigma(k)}} z \cdot \frac{\mathrm{d}}{\mathrm{d}z} x_{\sigma(k)}((z, \boldsymbol{b}_{-\sigma(k)}), \boldsymbol{p^{ctr}}) \mathrm{d}z \tag{16}$$

Substituting the allocation rule $x_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$ gives:

$$\begin{aligned}
p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) &= \int_0^{b_{\sigma(k)}} z \cdot \frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{ecpm_k}{\sum_{k'=1}^{q} ecpm_{k'}}\right) \mathrm{d}z \\
&= \int_0^{b_{\sigma(k)}} z \cdot \frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{z \cdot p^{ctr}_{\sigma(k)} p^{norm}_k}{w_{-k} + z \cdot p^{ctr}_{\sigma(k)} p^{norm}_k}\right) \mathrm{d}z \\
&= \int_0^{b_{\sigma(k)}} \frac{w_{-k} \cdot p^{ctr}_{\sigma(k)} p^{norm}_k \cdot z}{(w_{-k} + z \cdot p^{ctr}_{\sigma(k)} p^{norm}_k)^2} \mathrm{d}z \\
&= \int_0^{b_{\sigma(k)}} w_{-k} \cdot z \, \mathrm{d}\left(-\frac{1}{w_{-k} + z \cdot p^{ctr}_{\sigma(k)} p^{norm}_k}\right) \\
&= \int_0^{b_{\sigma(k)}} \frac{w_{-k}}{w_{-k} + p^{ctr}_{\sigma(k)} p^{norm}_k \cdot z} \mathrm{d}z - \left.\frac{w_{-k} \cdot z}{w_{-k} + p^{ctr}_{\sigma(k)} p^{norm}_k \cdot z}\right|_0^{b_{\sigma(k)}} \\
&= \left.\frac{w_{-k}}{p^{ctr}_{\sigma(k)} p^{norm}_k} \cdot \ln(w_{-k} + p^{ctr}_{\sigma(k)} p^{norm}_k z)\right|_0^{b_{\sigma(k)}} - \frac{w_{-k} \cdot b_{\sigma(k)}}{w_{-k} + ecpm_k} \\
&= \frac{w_{-k}}{p^{ctr}_{\sigma(k)} p^{norm}_k}[\ln(1+t) - \frac{t}{1+t}]
\end{aligned} \tag{17}$$

Since $\ln(1+t) - \frac{t}{1+t} \geq 0$ and $x \geq 0$, the advertiser's payment $p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) \geq 0$. The utility for advertiser $ad_{\sigma(k)}$ is given by:

$$\begin{aligned}
u_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) &= b_{\sigma(k)} \cdot \frac{ecpm_k}{\sum_{k'=1}^{q} ecpm_{k'}} - \frac{w_{-k}}{p^{ctr}_{\sigma(k)} p^{norm}_k} \\
&\left[\ln\left(\frac{w_{-k} + ecpm_k}{w_{-k}}\right) - \frac{ecpm_k}{w_{-k} + ecpm_k}\right] \\
&\geq b_{\sigma(k)} \cdot \frac{ecpm_k}{w_{-k} + ecpm_k} - \frac{w_{-k}}{p^{ctr}_{\sigma(k)} p^{norm}_k} \\
&\left[\frac{ecpm_k}{w_{-k}} - \frac{ecpm_k}{w_{-k} + ecpm_k}\right] \\
&= b_{\sigma(k)} \cdot \frac{ecpm_k}{w_{-k} + ecpm_k} - b_{\sigma(k)} + \frac{w_{-k} \cdot b_{\sigma(k)}}{w_{-k} + ecpm_k} \\
&= 0
\end{aligned} \tag{18}$$

Since $\ln(1+x) \leq x$, the utility of the advertiser is $u_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}}) \geq 0$, which means it is non-negative. Therefore, the auction mechanism $\mathcal{M}$ satisfies individual rationality.

■

V. Experiments

In this section, we present comprehensive experiments of the proposed MAE-AM using datasets[5]. The experimental code is implemented in Python, and conducted on an AMD Ryzen7 5800H with Radeon Graphics, 3.2GHz, and 16GB RAM. To ensure the reliability of the experimental results, each experiment is conducted five times under each configuration, and the mean values of all metrics are reported.

[4]Inspired by the literature [9], logarithmic social welfare is defined as $SW = \sum_{k=1}^{q} ecpm_k \cdot (\ln(Prom_k) + C)$, where $f(Prom_k)$ is a logarithmic function. The constant C is appropriately chosen to ensure that $f(Prom)$ remains non-negative.

[5]Additional experimental results, including those related to generated replies and prompts, are available on GitHub due to space constraints.

A. Experimental Setup

1) Datasets: We conduct large-scale experiments and evaluations using the following two Chinese and English datasets:

- **ADGEN** [19]: is a Chinese advertising dataset generated by models utilizing web tags and advertising copy. This dataset focuses on advertising content generation and primarily includes three product categories: shirts, skirts, and pants. Each sample contains detailed content descriptions of the associated product.
- **ATVI** [39]: is an English dataset comprising approximately 2,000 authentic business advertising scripts. It encompasses 26 different industries, including automotive, healthcare, and tourism, and others.

To the best of our knowledge, at the time of writing this paper, our work is the **first** to perform a comprehensive performance validation of the proposed method on large-scale datasets, which is different from the previous example verification experiments [8], [9], [21].

2) Implementation Details: For user queries, the datasets are first categorized by product or industry type. Subsequently, the classified advertisements are divided into 10 intervals based on content length. From each interval, 20 candidate advertisers are randomly sampled ($n = 20$), resulting in a diverse set of advertisements with varying lengths.

Next, leveraging advertising keywords from different categories, we employ a LLM to generate two types of query sets: recommendation-oriented and discussion-oriented queries. These queries facilitate the generation of both enumerative and integrated advertisements.

TABLE II: Experimental parameter table.

Parameters	Value Range and Default
p^{ctr}, p_k^{norm}	$p^{ctr} \in [0.3, 0.5], p_k^{norm} = (0.9)^k, k = 1, 2, \cdots, q$
q	$q \in \{1, 2, 3, 4, 5, 6\}$, default: $q = 4$
$L_{\max}$	$L_{\max} \in \{50, 100, 150, 200\}$, default: $L_{\max} = 150$
θ_1, θ_2, C	$\theta_1, \theta_2 \in [0, 1]$, default: $\theta_1 = 0.5, \theta_2 = 0.6, C = 3$
β	$\beta \in [0.2, 1.8]$

Unless otherwise specified, the experimental parameters settings are shown in TABLE II. Specifically, for the CTR settings, the synthetic dataset employs the base CTR $p^{ctr} \in [0.3, 0.5]$, referred to latest survey on advertising in Google and Microsoft[6]. For the real-world dataset, the actual observed CTR are utilized as the base CTR. The normalized CTR is defined as $p_k^{norm} = (0.9)^k$.

For the auction mechanism and reply generation settings, the following default configurations are adopted: the number of advertisement slots is $q = 4$, the maximum advertisement length is $L_{\max} = 150$, the parameters are configured as $\theta_1 = 0.5, \theta_2 = 0.6$, the function f is the logarithmic function with the parameter $C = 3$. The query type is set to recommendation-oriented, with enumerative advertisements adopted as generation format. The LLM community consists of six models: GPT-4 [3], Claude 3[7] [40], Moonshot-v1[8], GLM-4[9] [41], Ernie4.0 [42], Qwen [43], etc. Text similarity is computed using the pre-trained BERT [38] model, which evaluates semantic similarity between advertisement contents.

3) Compared methods: To evaluate the effectiveness of MAE-AM, we conduct comprehensive comparisons with both classical and latest methods. These comparisons are designed to assess the system's performance across diverse settings, showcasing its strengths relative to established approaches.

- **Greedy Auction (GA)**: allocates advertisements directly based on their content length, without leveraging LLM to optimize advertisement content during output.
- **Generalized Second-Price Auction (GSP)** [44]: employs the GSP mechanism and uses LLM to generates reply, assuming equal-length allocation of advertisements.
- **Sponsor Question Answer (SQA)** [21]: also utilizes the GSP, focusing on maximizing the platform's social welfare by placing a single LLM-optimized advertisement.
- **Auctions with LLM Summaries (ALS)** [8]: uses LLM to display multiple advertisements in a structured format and optimize the advertisement summary.
- **Ad Auctions for LLMs via RAG (AAL)** [9]: utilizes retrieval-augmented generation techniques to genetate integrated advertisements, presenting multiple relevant advertisements optimized through LLMs.

In pursuit of a uniform evaluative paradigm, we adopts: (1) the GA and GSP as foundational benchmarks for gauging the efficacy of auction mechanisms; (2) the remaining three benchmarks—SQA, ALS, and AAL—integrate the proposed auction mechanism into their respective frameworks. This integration is pivotal for enabling a direct and apples-to-apples comparison of various system design permutations.

This approach allows for a granular understanding of how different architectural choices can either enhance or detract from the performance of the auction mechanisms.

4) Metrics: We evaluate the performance of the proposed system using three auction-related metrics and three advertisement-related metrics as follows:

- **Social Welfare** (SW): the measure of the overall benefit to the system, computed by Equation (6).
- **Revenue** (Rev): the sum of the payment, formally defined as $Rev = \sum_{k=1}^{q} p_{\sigma(k)}(\boldsymbol{b}, \boldsymbol{p^{ctr}})$.
- **Utility** ($Util$): the difference between the valuation and payment for each advertiser, computed by Equation (1).
- **Advertiser Satisfaction** (S^a): the degree of satisfaction of the advertisers with the relevance of their contents.
- **User Satisfaction** (S^u): the degree of satisfaction of the users with the generated advertisement reply.
- **Score**: the multi-objective function for social welfare and satisfaction, computed by Algorithm 2.

[6]https://localiq.com/blog/search-advertising-benchmarks/

[7]https://www.anthropic.com/news/claude-3-family

[8]https://kimi.moonshot.cn/

[9]https://open.bigmodel.cn/

B. Results

1) Qualitative Analysis: We begin by presenting the qualitative results of the MAE-AM system on the two datasets and provide detailed case analyses. Taking the ATVI dataset as an example, the user query involves the description of the topic "travel", and is classified as a *recommendation query*. The system prompts the LLM to generate both an enumerative advertisement and integrated advertisement comprising four slots. An illustrative query is as follows (Detailed results for these advertisement formats are provided in the supplementary materials following the acknowledgment section):

> ***Query:*** *What travel agencies or booking platforms are popular for safe and enjoyable travel experiences? Please keep the reply briefly.*

According to the results, the MAE-AM system demonstrates its capability to generate highly relevant and contextually appropriate ad content for the given query, effectively integrating both enumerative and integrated advertisements within the base reply. This high degree of relevance in content design not only enhances the effectiveness of embedding but also improves user satisfaction.

Additionally, despite the inherent probability distributions of LLM-generated content, which does not guarantee fixed lengths, MAE-AM achieves robust control over text length through *meticulously designed prompt engineering*. The discrepancy between the generated text length and the target length remains minimal and falls within an acceptable range, ensuring the usability in practical applications.

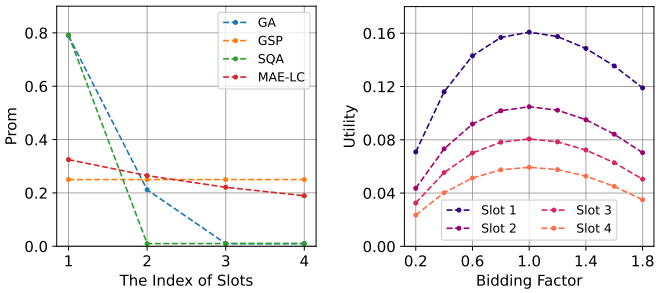
2) Overall Performance: We explore the system's performance improvements compared with aforementioned 5 methods under the default settings, which is summarized in TABLE III. According to the results, MAE-AM consistently outperformed other methods, particularly in social welfare, satisfaction, and overall score. Specifically, on the ADGEN, MAE-AM achieved average improvements of 19.30% and 28.54% in social welfare and overall score, respectively. Similarly, on the ATVI, these metrics improved by 25.53% and 35.75%, respectively.

Notably, while the GSP attained the highest revenue score, its underlying mechanism fails to satisfy DSIC and IR, leading to negative utility for advertisers — a result that is fundamentally unacceptable in practical settings. In contrast, MAE-AM not only prioritizes and maximizes social welfare but also achieves favorable outcomes in revenue and utility, demonstrating its practicality in real-world scenarios.

In terms of user satisfaction, MAE-AM significantly outperforms ALS and AAL. This improvement attributes to its effective integration of base reply in query-driven scenarios. By providing direct feedback to users, base replies play a pivotal role in enhancing satisfaction. In contrast, designs that focus solely on embedded advertisements often fail to comprehensively address user needs, leading to lower satisfaction.

While the SQA incorporates base replies, it restricts them to a single advertisement, which is insufficient for multi-slot advertisement. This limitation reduces advertiser satisfaction. In contrast, MAE-AM integrates base replies with multi-advertisement to meet the needs of both users and advertisers, thereby achieving higher satisfaction.

Fig. 4 illustrates the results of auction mechanisms in different slots, analyzed from two aspects: prominence allocation and DSIC property of the mechanism. As shown in Fig. 4(a), the prominence allocation in different slots indicates that GSP's even distribution fails to adequately distinguish advertisements. In contrast, SQA and GA exhibit excessive large allocation variance, leading to inefficient utilization of slots. By comparison, MAE-AM implements a strategy that allocates higher prominence to top-ranked slots, with prominence gradually diminishing for lower-ranked slots. This pattern aligns with the characteristics of PAM, ensuring a balance between effective resource utilization and maintaining necessary differentiation among advertisements.



(a) Prominence allocation in different slots.

(b) Utility achieved by advertisers under varying bidding factors.

Fig. 4: Results of the auction mechanism under different slots and bidding factors.

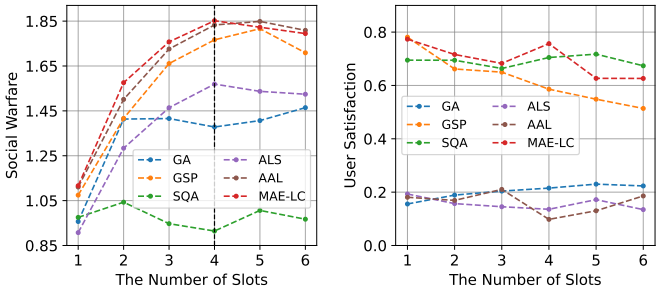
To verify the DSIC property of the proposed mechanism, we introduce a bidding factor defined as $\beta = \hat{b}_i / b_i, i \in [n]$, where $\hat{b}_i$ represents the *false bid* of advertiser ad_i. The experiment was conducted by fixing the bids of other advertisers while varying the bid of the current advertiser within the interval $\beta \in [0.2, 1.8]$. As shown in Fig. 4(b), the utility of the advertiser is maximized when the bidding factor $\beta = 1$, which corresponds to truthful bidding $\hat{b}_i = b_i$. This finding demonstrates that the proposed mechanism satisfies the DSIC property.

3) Effect of the number of slots: We analyze the impact of increasing the number of advertisement slots, content lengths, and candidate advertisers on system performance. Fig. 5 and 6 illustrates the relationship between the number of slots and four metrics, including social welfare, user satisfaction, advertisers satisfaction and score, under a fixed content length.

As shown in Fig. 5(a), with the increase in the number of slots, MAE-AM first increases, peaks at $q = 4$, and then decreases in terms of social welfare. This observation supports selecting $q = 4$ as the default setting for the subsequent experiments. These results suggest that an optimal number of slots enhances placement efficiency, whereas an excessive number reduces their prominence, therefore diminishing social welfare. By comparison, for the GA and SQA, the effect of

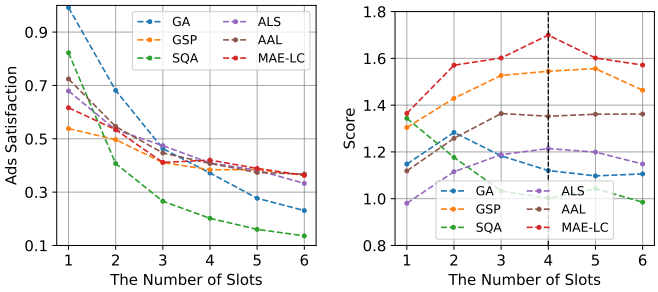
TABLE III: Performance comparison of different methods on the two datasets, detailed evaluation of social welfare, revenue, utility, satisfaction, and score. The best performances are indicated in **bold** font. Higher values for all metrics indicate better performance, and since the GSP may not satisfy IR, utility in the table can be *negative*.

Dataset	Metrics	GA	GSP	SQA	ALS	AAL	MAE-AM	Avg.
ADGEN	SW	1.3781	1.7665	0.9143	1.5693	1.8328	**1.8517**	1.5521
	Rev	0.2496	**1.0228**	0.3559	0.3178	0.3858	0.3710	0.4504
	$Util$	**0.6250**	−0.1793	0.3170	0.3906	0.4768	0.4587	0.3481
	S^a	0.3717	0.3839	0.2018	0.4093	0.4076	**0.4201**	0.3657
	S^u	0.2151	0.5859	0.7047	0.1357	0.0976	**0.7567**	0.4159
	$Score$	1.1203	1.5448	1.0019	1.2141	1.3524	**1.6994**	1.3221
ATVI	SW	1.1396	1.7662	1.1238	1.5446	1.8392	**1.9615**	1.5625
	Rev	0.3661	**1.0027**	0.3810	0.3171	0.3822	0.3907	0.4733
	$Util$	**0.5456**	−0.1818	0.5214	0.3890	0.4713	0.4887	0.3724
	S^a	0.2609	0.2797	0.1513	0.3102	0.3090	**0.3130**	0.2707
	S^u	0.2178	0.7124	0.7901	0.1652	0.1720	**0.9193**	0.4961
	$Score$	0.9231	1.5558	1.1450	1.1645	1.3440	**1.7931**	1.3209



(a) Social welfare. (b) User satisfaction.

Fig. 5: Illustration of social welfare & user satisfaction under varying number of slots.



(a) Ads satisfaction. (b) Overall score.

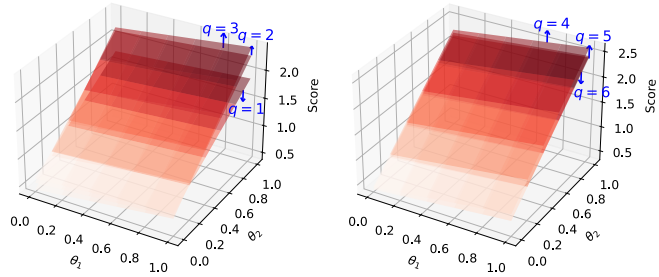
Fig. 6: Illustration of Ads satisfaction & overall score under varying number of slots.

increasing the number of slots on social welfare is relatively small, due to the single placement. These findings highlight the importance of design in rational slot quantity.

As shown in Fig. 5(b), with increase in the number of slots, methods have minimal effect on user satisfaction due to the constraint of a fixed maximum length. Notably, methods that include base replies significantly outperform those without them. These findings highlight that the inclusion of base replies effectively enhances the user experience.

As shown in Fig. 6(a), when the number of slots is limited ($q = 1$), both GA and SQA achieve relatively high levels of advertisers satisfaction. This can be attributed to the simplicity of GA, which avoids modifying advertisement content through LLMs, and SQA's embedding of only a single advertisement, ensuring high advertisers satisfaction for single placement. However, as the number of slots increases, advertiser satisfaction declines for both GA and SQA, while MAE-AM shows higher advertisers satisfaction. As shown in Fig. 6(b), MAE-AM consistently outperforms all other methods in terms of total score, regardless of the number of slots. Furthermore, it achieves highest score when $q = 4$ due to the social welfare.

4) Effect of parameter θ: In order to verify the effect of varying parameter settings for θ_1 and θ_2 on system performance, we compared the total scores under different slot numbers. As shown in Fig. 7(a), significant score improvements are observed for $q \in \{1, 2, 3\}$ as parameter increase.

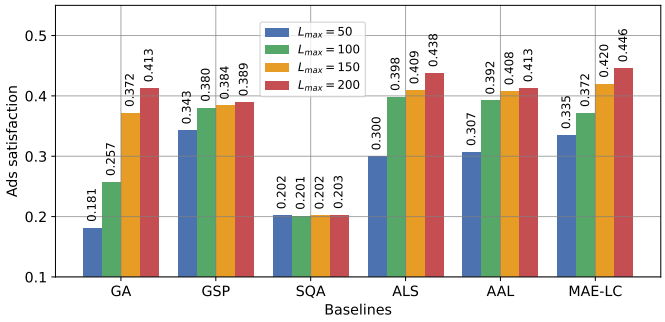


(a) $q \in \{1, 2, 3\}$. (b) $q \in \{4, 5, 6\}$.

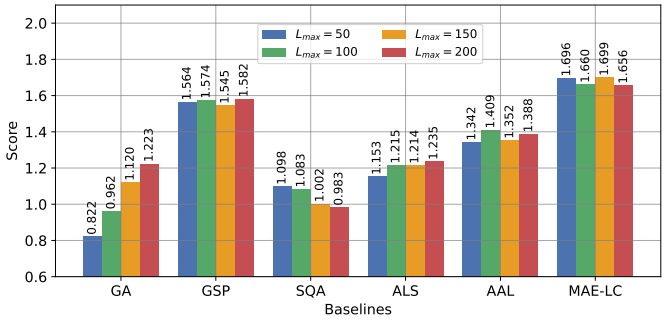
Fig. 7: Illustration of parameter settings on overall scores under different number of slots.

Fig. 7(b) shows that the overall score becomes less sensitive to parameter changes when $q \in \{4, 5, 6\}$, demonstrating a *more stable* trend. Therefore, selecting parameters requires consideration to balance satisfaction and social welfare.

5) Effect of different content lengths: To evaluate the effect of content length on system performance, we adjusted the maximum content length $L_{\max}$ while keeping the number of advertisement slots fixed at $q = 4$. As shown in Fig. 8(a), increasing the content length generally improves advertiser satisfaction, except for SQA. This improvement can be attributed to the additional space for information display, which enhances advertiser satisfaction. However, since SQA reply only contains a single advertiser, the increase in content length had a marginal impact on its satisfaction.



(a) Advertisers satisfaction.

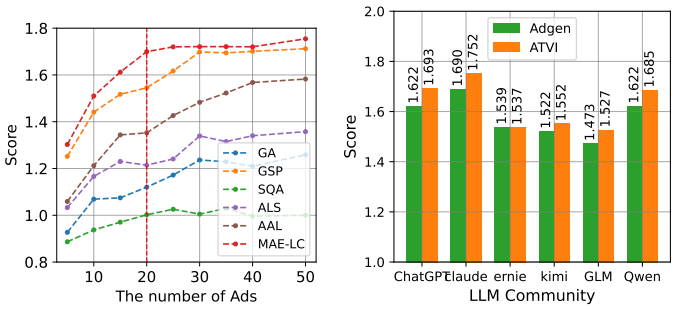


(b) Overall score.

Fig. 8: Illustration of ads satisfaction & overall score under varing ad content lengths.

Fig. 8(b) shows that the increase in content length had a relatively minor effect on the system's overall score, indicating that MAE-AM is insensitive to changes in content length.

6) Effect of the number of candidate advertisers and different LLMs: As shown in Fig. 9(a), we analyzed the effect of the number of candidate advertisers on the overall score. The experimental results show that the total score increases significantly as the number of candidate advertisers grows, primarily due to the availability of more contents with higher relevance to the query. However, when the number of candidates increases to a certain threshold, the total score tends to stabilize. It is worth noting that the "elbow" point of the candidates quantity is approximately 20 in our method. This is because, with a fixed number of slots, the inclusion of additional advertisers has a limited effect on improving the score, leading to diminishing returns. Therefore, it is essential to reasonably control the number of candidates to optimize system performance.



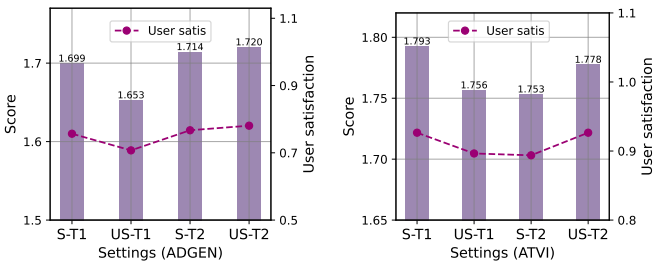
(a) Overall score under different number of candidate ads.

(b) Overall score of the two datasets under different LLMS.

Fig. 9: Overall score under different numbers of ads & LLMs.

As shown in Fig. 9(b), we analyzed the average performance of various LLMs within the LLM community, evaluated on the ADGEN and ATVI datasets. Among these models, Claude achieved the highest score. The results show that the models exhibit significant performance differences on both datasets, reflecting the uneven adaptation ability of LLMs in specific tasks. This indicates that the model selection should be based on the characteristics of the dataset and the task requirements to optimize overall performance in different application.

7) Effect of different types of generation formats: We analyzed the effect of different types of user queries and advertisement generation formats on system performance.



(a) Results of the ADGEN.

(b) Results of the ATVI.

Fig. 10: Total score (*left axis*) and user satisfaction (*right axis*) for the two datasets under different types of query and advertisement generation formats.

As shown in Fig. 10, we compared the effectiveness of two generating *enumerative advertisements* (S) and *integrated advertisements* (US) in terms of users satisfaction and total score under two types of queries: *recommendation-oriented* (T1) and *discussion-oriented* (T2). By combining these elements, we formed four types of question-answer formats, labeled as S-T1, US-T1, S-T2, and US-T2.

It is important to note that in enumerative advertisements, advertisement slot resemble the *positional structure* of search

engine advertisement. In contrast, integrated advertisements embed the content within cohesive paragraphs, with each advertisement slot corresponding to a sentence. Its index is *the order of sentences in paragraphs*. Unlike the rigid format of enumerative advertisements, integrated advertisements emphasize cohesive expression within the overall textual context.

The results indicate that for recommendation-oriented queries (T1), enumerative advertisements outperform integrated advertisements in both score and user satisfaction, likely due to their clear presentation of information. In contrast, for discussion-oriented queries (T2), integrated advertisements achieve higher scores and user satisfaction, attributed to their seamless integration into the contextual reply.

These findings align with intuitive expectations: enumerative advertisements are more suited for recommendation-oriented tasks, which enables users to quickly grasp key details. On the other hand, integrated advertisements are more effective for tasks requiring textual coherence within discussion-oriented contexts. These results highlight the importance of tailoring formats to user query to optimize both user satisfaction and overall system performance.

VI. Conclusion

In this paper, we proposed the MAE-AM system, a query-driven multi-advertisement embedding and auction mechanism in LLM. The system first adopted a Top-q greedy multi-advertisement auction mechanism to maximize social welfare. Next, an optimal reply generation scheme based on multi-objective coordination was utilized to generate candidate embedded advertisements replies through prompt engineering within the LLM community. Furthermore, we provide a theoretical proof to demonstrate that the mechanism satisfies properties such as DSIC and IR. Empirical results showed that MAE-AM improved improves both user and advertiser satisfaction while maximizing social welfare.

Currently, the proposed system is confined to generating natural language text advertisements and does not included in other multimodal forms, such as image, audio and video advertisements. Consequently, our future research will focus on developing systems for multimodal advertisement generation that integrate user preferences, enabling dynamic adjustments during the generation process.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 61872313. We also extend our gratitude to the anonymous reviewers for their valuable feedback on an earlier draft of this paper.

Supplementary Materials

The results are provided of the two advertisement formats discussed in section V-B1: enumerative and integrated advertisement. Detailed results for these formats are outlined below.

Here's a detailed example results of a enumerative advertisement, which includes listed advertisements:

> *Popular and reputable travel agencies and booking platforms include Expedia, Travelocity, Orbitz, and TripAdvisor. They offer a wide range of travel options, competitive prices, and customer reviews to help ensure safe and enjoyable travel experiences.*
> ***1. Renaissance Hotels:*** *Experience extraordinary hotels in extraordinary places with Renaissance Hotels. Each location offers unique designs, world-class amenities, and personalized service tailored to your needs. Whether you're traveling for business or leisure, let us elevate your stay and create unforgettable memories in stunning destinations around the globe.*
> ***2. Northport Inn:*** *Indulge in the delightful flavors of Northport Inn, where two luscious varieties meet in one great soft-baked crust. Perfectly crafted for a satisfying treat, our irresistible offerings are sure to please your taste buds and make every moment enjoyable.*
> ***3. Breakstone's Cottage Doubles:*** *Enjoy the delicious taste of Breakstone's Cottage Doubles, where we've reduced the sugar but maintained the great protein. Perfect for a nutritious snack or part of a balanced meal, our cottage cheese delivers flavor and health in every bite.*
> ***4. Disney Cruise Line:*** *Escape on a Disney Cruise Line adventure, where fantasy meets reality. Discover enchanting destinations, exceptional service, and unforgettable experiences tailored for families and individuals alike, creating magical memories that last a lifetime.*

Here's a detailed example results of a integrated advertisement, blending the base reply with advertisements:

> *Popular and reputable travel agencies and booking platforms include Expedia, Travelocity, Orbitz, and TripAdvisor. They offer a wide range of travel options, competitive prices, and customer reviews to help ensure safe and enjoyable travel experiences. Experience extraordinary hotels in extraordinary places with* ***Renaissance Hotels****. Each location offers unique designs, world-class amenities, and personalized service tailored to your needs. Whether you're traveling for business or leisure, let us elevate your stay and create unforgettable memories in stunning destinations around the globe. Indulge in the delightful flavors of* ***Northport Inn****, where two luscious varieties meet in one great soft-baked crust. Perfectly crafted for a satisfying treat, our irresistible offerings are sure to please your taste buds and make every moment enjoyable. Enjoy the delicious taste of* ***Breakstone's Cottage Doubles****, where we've reduced the sugar but maintained the great protein. Perfect for a nutritious snack or part of a balanced meal, our cottage cheese delivers flavor and health in every bite. Escape on a* ***Disney Cruise Line*** *adventure, where fantasy meets reality. Discover enchanting destinations, exceptional service, and unforgettable experiences tailored for families and individuals alike, creating magical memories that last a lifetime.*

REFERENCES

[1] D. Chaffey and F. Ellis-Chadwick, *Digital marketing*. Pearson uk, 2019.

[2] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[4] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez *et al.*, "Code llama: Open foundation models for code," *arXiv preprint arXiv:2308.12950*, 2023.

[5] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 81–87.

[7] Y. Juan, D. Lefortier, and O. Chapelle, "Field-aware factorization machines in a real-world online advertising system," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 680–688.

[8] A. Dubey, Z. Feng, R. Kidambi, A. Mehta, and D. Wang, "Auctions with llm summaries," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 713–722.

[9] M. Hajiaghayi, S. Lahaie, and K. Rezaei, "Ad auctions for llms via retrieval augmented generation," *arXiv preprint arXiv:2406.09459*, 2024.

[10] H. R. Varian, "Position auctions," *international Journal of industrial Organization*, vol. 25, no. 6, pp. 1163–1178, 2007.

[11] A. Ghose and S. Yang, "An empirical analysis of search engine advertising: Sponsored search in electronic markets," *Management science*, vol. 55, no. 10, pp. 1605–1622, 2009.

[12] A. Goldfarb and C. Tucker, "Search engine advertising: Channel substitution when pricing ads to context," *Management Science*, vol. 57, no. 3, pp. 458–470, 2011.

[13] C. Borgs, J. Chayes, N. Immorlica, K. Jain, O. Etesami, and M. Mahdian, "Dynamics of bid optimization in online advertisement auctions," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 531–540.

[14] R. Fang and X. Li, "Advertising a second-price auction," *Journal of Mathematical Economics*, vol. 61, pp. 246–252, 2015.

[15] S. Feizi, M. Hajiaghayi, K. Rezaei, and S. Shin, "Online advertisements with llms: Opportunities and challenges," *arXiv preprint arXiv:2311.07601*, 2023.

[16] L. Liu, J. Meng, and Y. Yang, "Llm technologies and information search," *Journal of Economy and Technology*, vol. 2, pp. 269–277, 2024.

[17] T.-M. Choi, X. Li, and C. Ma, "Search-based advertising auctions with choice-based budget constraint," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 8, pp. 1178–1186, 2015.

[18] T. Payne, E. David, N. R. Jennings, and M. Sharifi, "Auction mechanisms for efficient advertisement selection on public displays," in *ECAI*, 2006, pp. 285–289.

[19] S. Zhihong, H. Minlie, W. Jiangtao, W. Xu, and Z. Xiaoyan, "Long and diverse text generation with planning-based hierarchical variational model," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 2019, pp. 3255–3266.

[20] Y. Wang, Z. Sha, K. Lin, C. Feng, K. Zhu, L. Wang, X. Jiao, F. Huang, C. Ye, D. He *et al.*, "One-step reach: Llm-based keyword generation for sponsored search advertising," in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 1604–1608.

[21] T. Mordo, M. Tennenholtz, and O. Kurland, "Sponsored question answering," in *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, 2024, pp. 167–173.

[22] S. Hu, T. Huang, F. Ilhan, S. Tekin, G. Liu, R. Kompella, and L. Liu, "A survey on large language model-based game agents," *arXiv preprint arXiv:2404.02039*, 2024.

[23] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large language models and games: A survey and roadmap," *arXiv preprint arXiv:2402.18659*, 2024.

[24] X. Xu, Y. Wang, C. Xu, Z. Ding, J. Jiang, Z. Ding, and B. F. Karlsson, "A survey on game playing agents and large models: Methods, applications, and challenges," *arXiv preprint arXiv:2403.10249*, 2024.

[25] B. Lamichhane, J. Palardy, and A. K. Singh, "The nuances of large-language-model-agent performance in simple english auctions," *Empirical Economics Letters*, vol. 22, no. 1, 2023.

[26] N. Immorlica, B. Lucier, and A. Slivkins, "Generative ai as economic agents," *ACM SIGecom Exchanges*, vol. 22, no. 1, pp. 93–109, 2024.

[27] Y. Liu, P. Sun, and H. Li, "Large language models as agents in two-player games," *arXiv preprint arXiv:2402.08078*, 2024.

[28] M. Chahine, T.-H. Wang, H. Zhang, W. Xiao, D. Rus, and C. Gan, "Large language models can design game-theoretic objectives for multi-agent planning."

[29] P. Duetting, V. Mirrokni, R. Paes Leme, H. Xu, and S. Zuo, "Mechanism design for large language models," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 144–155.

[30] H. Sun, Y. Chen, S. Wang, W. Chen, and X. Deng, "Mechanism design for llm fine-tuning with multiple reward models," *arXiv preprint arXiv:2405.16276*, 2024.

[31] J. Xu, Z. Zhang, Z. Lu, X. Deng, M. P. Wellman, C. Yu, S. Dou, Y. Huo, Z. Xu, Z. Duan *et al.*, "Auto-bidding in large-scale auctions: Learning decision-making in uncertain and competitive games," in *NeurIPS 2024 Competition Track*.

[32] H. Zhang, L. Niu, Z. Zheng, Z. Zhang, S. Gu, F. Wu, C. Yu, J. Xu, G. Chen, and B. Zheng, "A personalized automated bidding framework for fairness-aware online advertising," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5544–5553.

[33] I. Zelch, M. Hagen, and M. Potthast, "A user study on the acceptance of native advertising in generative ir," in *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, 2024, pp. 142–152.

[34] B. J. Tang, K. Sun, N. T. Curran, F. Schaub, and K. G. Shin, "Genai advertising: Risks of personalizing ads with llms," *arXiv preprint arXiv:2409.15436*, 2024.

[35] E. Meguellati, L. Han, A. Bernstein, S. Sadiq, and G. Demartini, "How good are llms in generating personalized advertisements?" in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 826–829.

[36] J. Sun, T. Zhang, H. Jiang, K. Huang, C. Luo, J. Wu, J. Wu, A. Zhang, and X. Wang, "Large language models empower personalized valuation in auction," *arXiv preprint arXiv:2410.15817*, 2024.

[37] R. B. Myerson, "Optimal auction design," *Mathematics of operations research*, vol. 6, no. 1, pp. 58–73, 1981.

[38] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[39] KevinHartman, "Advertisement transcripts from various industries," https://www.kaggle.com/datasets/kevinhartman0/advertisement-transcripts-from-various-industries/data, 2021.

[40] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[41] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[42] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu *et al.*, "Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv:2107.02137*, 2021.

[43] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, 2024.

[44] B. Edelman and M. Ostrovsky, "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," *American economic review*, vol. 97, no. 1, pp. 242–259, 2007.