Strategic Prompt Pricing for AIGC Services: A User-Centric Approach

Xiang Li, Bing Luo, Jianwei Huang, Yuan Luo

Abstract—**The rapid growth of AI-generated content (AIGC) services has created an urgent need for effective prompt pricing strategies, yet current approaches overlook users' strategic two-step decision-making process in selecting and utilizing generative AI models. This oversight creates two key technical challenges: quantifying the relationship between user prompt capabilities and generation outcomes, and optimizing platform payoff while accounting for heterogeneous user behaviors. We address these challenges by introducing prompt ambiguity, a theoretical framework that captures users' varying abilities in prompt engineering, and developing an Optimal Prompt Pricing (OPP) algorithm. Our analysis reveals a counterintuitive insight: users with higher prompt ambiguity (i.e., lower capability) exhibit non-monotonic prompt usage patterns, first increasing then decreasing with ambiguity levels, reflecting complex changes in marginal utility. Experimental evaluation using a character-level GPT-like model demonstrates that our OPP algorithm achieves up to 31.72% improvement in platform payoff compared to existing pricing mechanisms, validating the importance of user-centric prompt pricing in AIGC services.**

I. Introduction

With impressive progress in generative artificial intelligence (GAI) [1], AI-generated content (AIGC) services are growing in popularity, evidenced by platforms such as Monica [2] and Poe [3]. Figure 1 illustrates a typical framework of the AIGC service process, where the platform usually integrates multiple GAI models, e.g., GPT-4 [4] and GPT-4o [5], each with different performance of generating content in response to user prompts. Users will purchase AIGC services from the platform under a usage metric, i.e., the number of prompts [6], to interact with distinct models and obtain desired content.

This novel AI service paradigm has promoted the emergence of related research [7]–[12], as summarized in surveys [13]–[15]. For example, Du *et al.* in [7] designed a dynamic service selection scheme to enhance AIGC quality. Dütting *et al.* in [8] proposed an AIGC advertising auction mechanism. Liu *et al.* in [12] incorporated semantic communication to achieve efficient content generation. However, these studies primarily address AIGC service optimization, leaving a significant research gap in platform prompt pricing design.

Xiang Li is with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), and the School of Science and Engineering (SSE), The Chinese University of Hong Kong, Shenzhen (E-mail: xiangli2@link.cuhk.edu.cn). Bing Luo is with Data Science Research Center, Duke Kunshan University, Jiangsu (E-mail: bl291@duke.edu). Jianwei Huang is with SSE, AIRS, Shenzhen Key Laboratory of Crowd Intelligence Empowered Low-Carbon Energy Network, and CSIJRI Joint Research Centre on Smart Energy Storage, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China (Corresponding Author, E-mail: jianweihuang@cuhk.edu.cn). Yuan Luo is with SSE, AIRS, and Shenzhen Key Laboratory of Crowd Intelligence Empowered Low-Carbon Energy Network, The Chinese University of Hong Kong, Shenzhen (E-mail: luoyuan@cuhk.edu.cn).

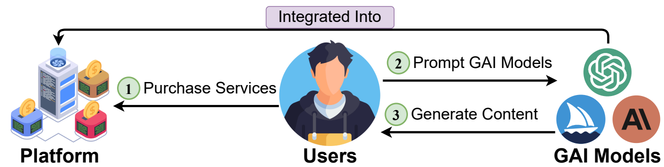


Fig. 1. Framework of AIGC Service Process.

Current prompt pricing (e.g., [2], [3], [16]) typically relies on GAI models' cost or performance metric, neglecting users' strategic decision-making process. In AIGC services, the platform decides the prompt price of each model and necessitates users to make two-step utilization decisions: which model to utilize and how to utilize it (i.e., the number of prompts for interaction [6], [15]). Users will decide both the optimal model and prompt number to maximize individual payoff, based on their varying prompt capabilities [17]. Failing to account for such heterogeneous decisions in prompt pricing could not only reduce the platform's payoff, but also discourage users (e.g., those with low prompt capability) from AIGC services. Thus, to design the optimal prompt pricing mechanism, we first delve into users' strategic decisions in AIGC services, which motivates our key question as follows:

Key Question 1. *What is the optimal utilization strategy for users with different prompt capabilities in AIGC services?*

The prompt-based content generation of AIGC services poses unique challenges to user decision-making. Specifically, the user's utility derived from generated content depends on both model performance and user prompts (i.e., prompt number and quality, as highlighted in [17], [18]). This correlation and inherent uncertainty [19] make it difficult to assess each prompt's value across diverse models and users, complicating decisions on selecting and prompting GAI models.

To address the above challenges from interactions between models and prompts, we introduce the concept of prompt ambiguity, inspired by in-context learning [20]–[22]. This mathematical framework quantifies user prompts as the probability of correctly conveying the intended task, enabling us to derive optimal utilization decisions for heterogeneous users. With users' optimal strategy in AIGC services, we can study our second key question of this paper:

Key Question 2. *What is the platform's optimal prompt pricing mechanism for AIGC services?*

ISBN 978-3-903176-73-7 © 2025 IFIP

The design of optimal prompt pricing presents a challenging bi-level non-convex optimization problem, due to the heterogeneity of users' prompt capabilities and strategic decisions. The platform must maximize payoff while accounting for users' two-step decision-making process, forming a complex bi-level structure. Prompt prices across various GAI models are also ineluctably intertwined with the trade-off between user strategies in selecting and prompting the model, resulting in the non-convexity of this mechanism design problem.

Against this background, we first transform the problem into an unconstrained piecewise form to determine optimal prompt pricing for users with consistent prompt capabilities (homogeneous case). Then, by exploring the relationship between user utilization decisions and model prompt prices in this simplified scenario, we establish the price upper bound for the more general heterogeneous user case. According to these insights, we propose an Optimal Prompt Pricing (OPP) algorithm, which decomposes prompt pricing design into tractable sub-problems for platform payoff maximization.

The key results and contributions of the paper are as follows:

- *Optimal Prompt Pricing Framework:* We address the fundamental challenge of modeling strategic interactions between the platform and heterogeneous users in AIGC services. Our framework transforms this bi-level optimization problem into tractable sub-problems through theoretical bounds analysis, providing the first systematic solution for prompt-based pricing mechanisms.
- *User Utilization Strategy Analysis:* We tackle the challenge of quantifying diverse user prompt capabilities in AIGC services. By introducing prompt ambiguity based on in-context learning [20]–[22], we establish a mathematical framework that precisely captures how users select and interact with different GAI models.
- *Impact of User Prompt Ambiguity:* We solve the non-intuitive relationship between user prompt capability and optimal prompt usage. Our analysis reveals that higher prompt ambiguity (lower capability) creates a non-monotonic pattern in prompt number: first increases and then decreases with ambiguity levels, corresponding to variations in user marginal utility of additional prompts.
- *Experimental Evaluation:* We bridge the gap between theoretical analysis and practical implementation using a character-level GPT-like model [23] and synthetic language data set [20]–[22]. The extensive experiments validate our analysis and demonstrate that our proposed OPP algorithm can improve the platform's payoff by up to 31.72%, compared to existing pricing mechanisms.

The remainder of this paper is organized as follows. Section II presents the system model and problem formulation in AIGC services. Section III and Section IV respectively derive heterogeneous users' optimal utilization strategies and the platform's optimal prompt pricing. We provide experimental evaluation in Section V and conclude the paper in Section VI. Due to the page limit, we leave detailed proofs of our results and additional theoretical analysis in the online appendix [24].

II. System Model and Problem Formulation

In this section, we first introduce the AIGC service process, a two-stage Stackelberg game between users and the platform. Then, we elaborated on modeling heterogeneous users and the platform's optimal prompt pricing mechanism design problem.

A. The AIGC Service Process

We consider a typical two-stage process of AIGC services between heterogeneous users and a platform, outlined below:

- **Stage 1:** *Platform* publishes a set of GAI models, each with a specific prompt price and performance characteristic of generating content in response to user prompts.
- **Stage 2:** *Users* make two-step utilization decisions for AIGC services, determining which GAI model to utilize and the number of prompts required for its utilization.

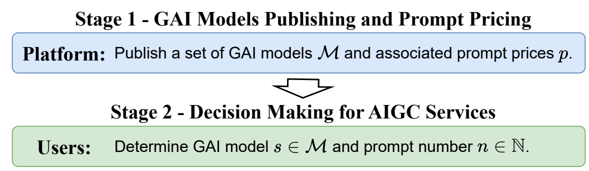


Fig. 2. Two-stage Stackelberg Game of AIGC Services.

Next, we present each stage in detail, depicted in Figure 2.

1) Stage 1: The platform offers a set of GAI models with different performances and decides each model's prompt price.

- *Published GAI Model Set:* The platform publishes a set of GAI models $\mathcal{M} = \{M_L, M_H\}$ with distinct performance tiers: a low-performance model M_L and a high-performance model M_H. This binary categorization, prevalent in existing platforms (e.g., [2], [3], [16]), enables focused theoretical analysis while maintaining generalizability to multi-model scenarios.
- *Prompt Prices:* The platform establishes prompt prices $\boldsymbol{p} = \{p_m, \forall m \in \mathcal{M}\}$, where p_m denotes the price per prompt for model $m \in \mathcal{M}$. This usage-based pricing scheme aligns with current industry practices. For example, Poe [3] charges approximately $\$0.008$ per prompt for GPT-4 [4] and $\$0.01$ for GPT-4o [5].
- *Operational Costs:* Each prompt incurs operational costs for the platform, represented by $\mathcal{C} = \{C_m, \forall m \in \mathcal{M}\}$, where C_m denotes the per-prompt cost for GAI model $m \in \mathcal{M}$. These costs encompass both computation [6] and communication [15] resources. For reference, GPT-4's operational cost is about $\$0.0036$ per prompt [25].

2) Stage 2: Users decide both which GAI model to utilize and the number of prompts required for its utilization.

- *Model Selection:* Users choose a specific model $s \in \mathcal{M}$ from the available options based on the platform's published model set $\mathcal{M}$ and pricing strategy $\boldsymbol{p}$.
- *Prompt Strategy:* Users determine the optimal number of prompts $n \in \mathbb{N}$ needed for their task. This decision depends on both the selected model's performance and the user's prompt capability.

These decisions in Stage 2 are interdependent, and their optimal values are analyzed in detail in the following section.

B. Modeling of Heterogeneous Users

This subsection characterizes the user's prompt behaviors, expected utility, and expected payoff in AIGC services.

1) User's Prompt Behaviors: Users interact with GAI models by conveying a specific intention (task) through prompts to generate desired content, such as instructions like "avoid long sentences" or "highlight key messages" when polishing text (see Figure 3). To formalize this interaction, we establish a mathematical framework where Γ denotes a discrete and complete space of user intentions, with each intention $\gamma \in \Gamma$ being unique, and $\mathcal{X}_\gamma$ represents the set of user prompts associated with intention γ. Each prompt $x \in \mathcal{X}_\gamma$ is independently[1] generated from a distribution [26], [27] conditioned on the corresponding intention γ, providing a foundational structure for analyzing user-model interactions in AIGC services.

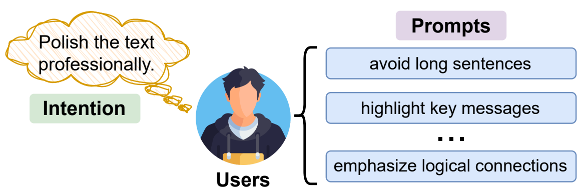


Fig. 3. Example of User Prompts and Intention.

In practice, users' varying prompt capabilities often result in ambiguous communication of their target intention $\gamma \in \Gamma$. Drawing from in-context learning literature [20]–[22], we introduce the concept of prompt ambiguity to quantify these heterogeneous prompt capabilities:

Definition 1 (Prompt Ambiguity)**.** *A user prompt $x \in \mathcal{X}_\gamma$ is ϵ-ambiguous*[2] *for intention $\gamma \in \Gamma$ with $\epsilon \in (0,1)$, if*

$$Prob\,(\gamma|x) \geq 1-\epsilon, \tag{1}$$

which indicates that the probability of correctly inferring the target intention γ from user prompt x is at least $1-\epsilon$.

Building upon Definition 1 and existing literature [20]–[22], we establish the relationship between user prompts and target intention in Lemma 1.

Lemma 1. *For a user with prompt ambiguity ϵ, the probability of inferring target intention $\gamma \in \Gamma$ from $k \in \mathbb{N}^+$ prompts is:*

$$Prob(\gamma_1 = \gamma, \cdots, \gamma_k = \gamma | x_1, \cdots, x_k) \geq 1 - \epsilon^k, \tag{2}$$

where x_k denotes the user's k-th prompt and γ_k represents the intention conveyed by prompt x_k, for any $k \in \mathbb{N}^+$.

The detailed proof of Lemma 1 is provided in the appendix [24]. To capture the diversity in users' prompt capabilities, we employ a probability density function $f(\epsilon)$ to characterize the distribution of prompt ambiguity across the user population.

2) User's Expected Utility: Within intention space Γ, a user's utility from generated content is determined by two factors: the selected model's performance and the probability of successfully conveying the target intention through prompts. For any published model $m \in \mathcal{M}$, we define U_m as the utility that maps from intention space to generated content, indicating the model's performance in responding to user intentions (e.g., prediction accuracy in semantic tasks [20]). Accordingly, the utility mapping of the user's selected model $s \in \mathcal{M}$ is U_s. Based on Lemma 1, for a user's GAI model selection $s \in \mathcal{M}$, prompt number $n \in \mathbb{N}$, and prompt ambiguity $\epsilon \in (0,1)$, the user's expected utility is expressed as:

$$(1-\epsilon^n) \cdot U_s, \tag{3}$$

where U_s represents the utility of model $s \in \mathcal{M}$.

3) User's Expected Payoff: The user's payoff π_{user} is defined as the difference between expected utility $(1-\epsilon^n)U_s$ and the total prompting price np_s:

$$\pi_{\text{user}}(s, n, \boldsymbol{p}, \epsilon) = (1-\epsilon^n) \cdot U_s - n \cdot p_s. \tag{4}$$

Users make optimal decisions by selecting model choice $s^* \in \mathcal{M}$ and prompt number $n^* \in \mathbb{N}$ to maximize payoff π_{user} in (4). These decisions are based on prompt prices $\boldsymbol{p}$ and their individual prompt ambiguity $\epsilon \in (0,1)$. To emphasize the interdependence between optimal decisions, we denote them as $s^*(n^*, \boldsymbol{p})$ and $n^*(s, p_s, \epsilon)$, respectively. This formalization provides the foundation for analyzing the platform's prompt pricing mechanism design problem.

C. Platform's Prompt Pricing Mechanism Design Problem

In AIGC services, the platform decides the optimal prompt pricing $\boldsymbol{p}^* = \{p_m^*, \forall m \in \mathcal{M}\}$ to maximize its payoff across all GAI models and users. We first summarize the platform's expected payoff $\pi_{\text{platform}}(\boldsymbol{p})$ and then formally introduce its problem of prompt pricing mechanism design.

For the platform, its expected payoff $\pi_{\text{platform}}(\boldsymbol{p})$ is the aggregate of that from each GAI model. Specifically, given any model $m \in \mathcal{M}$, the platform's per-prompt payoff is $p_m - C_m$, where p_m and C_m, respectively, represent the model's prompt price and prompt cost. Through the prompt ambiguity density function $f(\epsilon)$, we can further determine the expected number of user prompts for model $m \in \mathcal{M}$:

$$N_m(\boldsymbol{p}) = \int_0^1 f(\epsilon) \cdot n^*(m, p_m, \epsilon) \cdot \mathbb{1}_{s^*(n^*, \boldsymbol{p}) = m} \, d\epsilon, \tag{5}$$

where indicator function $\mathbb{1}_{s^*(n^*, \boldsymbol{p}) = m}$ indicates that the optimal GAI model for the user is $m \in \mathcal{M}$. Accordingly, the expected payoff of the platform in AIGC services is:

$$\pi_{\text{platform}}(\boldsymbol{p}) = \sum\nolimits_{m \in \mathcal{M}} (p_m - C_m) \cdot N_m(\boldsymbol{p}). \tag{6}$$

To maximize payoff $\pi_{\text{platform}}(\boldsymbol{p})$ in (6), the platform needs to account for each user's self-interested decisions $s^*(n^*, \boldsymbol{p})$ and $n^*(s, p_s, \epsilon)$. By incorporating heterogeneous users' strategic

[1] This initial work focuses on scenarios where user prompts are independent of each other. For dependent cases, e.g., chain-of-thought prompting [18], the analysis is similar but involves more complicated prompt correlations.

[2] To avoid trivial cases, we assume $\epsilon \in (0,1)$. If $\epsilon = 0$, the user always requires only one prompt. Conversely, when $\epsilon = 1$, the prompt is meaningless.

decisions, we formulate the platform's optimal prompt pricing problem as a bi-level optimization in Problem 1.

> **Problem 1** (Prompt Pricing Mechanism Design)**.**
>
> $$\max_{\boldsymbol{p}} \sum\nolimits_{m\in\mathcal{M}} (p_m - C_m)\cdot N_m(\boldsymbol{p}),$$
> $$s.t.\ \pi_{user}(s,n,\boldsymbol{p},\epsilon) = (1-\epsilon^n)\cdot U_s - n\cdot p_s,$$
> $$N_m(\boldsymbol{p}) = \int_0^1 f(\epsilon)\cdot n^*(m,p_m,\epsilon)\cdot \mathbb{1}_{s^*(n^*,\boldsymbol{p})=m}\, d\epsilon,$$
> $$n^*(s,p_s,\epsilon) = \underset{n\in\mathbb{N}}{\operatorname{argmax}}\ \pi_{user}(s,n,\boldsymbol{p},\epsilon),\ \forall s\in\mathcal{M},$$
> $$s^*(n^*,\boldsymbol{p}) = \underset{s\in\mathcal{M}}{\operatorname{argmax}}\ \pi_{user}(s,n^*(s,p_s,\epsilon),\boldsymbol{p},\epsilon),$$
> $$var.\ \boldsymbol{p} = \{p_m, \forall m\in\mathcal{M}\}.$$

Having established the system model and problem formulation for AIGC services, we tackle this two-stage Stackelberg game using backward induction [28]. Our study follows a sequential approach: Section III examines users' strategic utilization decisions in Stage 2, and subsequently, Section IV leverages these insights to derive the platform's optimal prompt pricing mechanism in Stage 1.

III. Heterogeneous Users' Utilization Strategy

This section analyzes heterogeneous users' optimal utilization strategies in AIGC services. Our analysis progresses from single-user optimization to multiple-user scenarios, revealing how prompt ambiguity and pricing affect users' strategic behaviors and expected payoff.

A. The User's Optimal Utilization Strategy

According to the published model set $\mathcal{M}$ and prompt pricing $\boldsymbol{p} = \{p_m, \forall m\in\mathcal{M}\}$, each user maximizes individual payoff $\pi_{\text{user}}(s,n,\boldsymbol{p},\epsilon)$ by deciding the optimal model $s^*(n^*,\boldsymbol{p})$ and prompt number $n^*(s,p_s,\epsilon)$. Given model decision $s\in\mathcal{M}$ and ambiguity $\epsilon\in(0,1)$, Theorem 1 derives the optimal prompt strategy $n^*(s,p_s,\epsilon)$ for the user.

Theorem 1. *For any selected model $s\in\mathcal{M}$ with utility mapping U_s and prompt price p_s, a user with prompt ambiguity $\epsilon\in(0,1)$ has the following optimal strategy:*

$$n^*(s,p_s,\epsilon) = \begin{cases} \lfloor \log_\epsilon \dfrac{\epsilon p_s}{(1-\epsilon)U_s} \rfloor, \textit{if } 0\le p_s \le (1-\epsilon)U_s, & \text{(7a)} \\ 0, \qquad\qquad\qquad \textit{if } p_s > (1-\epsilon)U_s, & \text{(7b)} \end{cases}$$

where $\lfloor\cdot\rfloor$ is the floor function.

We give the proof of Theorem 1 in online appendix [24]. Theorem 1 provides a theoretical foundation for understanding the relationship between optimal prompt number $n^*(s,p_s,\epsilon)$ and two key factors: prompt price p_s and utility mapping U_s. For any selected model $s\in\mathcal{M}$, utility mapping U_s establishes a price ceiling of $(1-\epsilon)U_s$ that a user with ambiguity ϵ will accept. The optimal number of prompts will increase with U_s and decrease with prompt price p_s.

Following Theorem 1, we can conclude the user's optimal model selection $s^*(n^*,\boldsymbol{p})$ from model set $\mathcal{M}$:

$$\begin{aligned} s^*(n^*,\boldsymbol{p}) = \operatorname{argmax}_{s\in\mathcal{M}}(1-\epsilon\hat{}\lfloor \log_\epsilon \frac{\epsilon p_s}{(1-\epsilon)U_s}\rfloor)\cdot U_s \\ - \lfloor \log_\epsilon \frac{\epsilon p_s}{(1-\epsilon)U_s}\rfloor \cdot p_s. \end{aligned} \tag{8}$$

This formulation captures the balance between utility maximization and price minimization in model selection.

B. Multiple Users with Heterogeneous Prompt Ambiguities

After studying single user's utilization decisions $s^*(n^*,\boldsymbol{p})$ and $n^*(s,p_s,\epsilon)$, we proceed to extend our analysis to encompass multiple users with heterogeneous prompt ambiguities. In particular, we start from the upper bound of heterogeneous users' optimal prompt number $\bar{n}^*(m)\in\mathbb{N}^+$ in Lemma 2, in order to investigate the impact of ambiguity ϵ on users' optimal strategies. This will form the foundation for designing the optimal prompt pricing mechanism in Stage 1.

Lemma 2. *For users with varying ambiguity $\epsilon\in(0,1)$, the upper bound $\bar{n}^*(m)$ on users' optimal prompt number to GAI model $m\in\mathcal{M}$ is:*

$$\bar{n}^*(m) = \min\left\{k\in\mathbb{N}^+ : \frac{k^k}{(k+1)^{k+1}} < \frac{p_m}{U_m}\right\}. \tag{9}$$

Lemma 2 identifies the critical ratio p_m/U_m that affects the upper bound $\bar{n}^*(m)$ of heterogeneous users' optimal prompt number for model $m\in\mathcal{M}$. Building on this, Proposition 1 elaborates on the relationship between prompt ambiguity $\epsilon\in(0,1)$ and the optimal prompt strategy $n^*(s,p_s,\epsilon)$ of users.

Proposition 1. *Given model decision $s\in\mathcal{M}$, utility mapping U_s and prompt price p_s, users' optimal prompt strategy $n^*(s,p_s,\epsilon)$ varies with ambiguity $\epsilon\in(0,1)$ as follows:*

- *if $p_s = 0$: $n^*(s,p_s,\epsilon) = \infty$.*
- *if $p_s \in (0, U_s/4]$: $n^*(s,p_s,\epsilon)$ first increases and then decreases with ϵ.*
- *if $p_s \in (U_s/4, U_s)$: $n^*(s,p_s,\epsilon)$ decreases with ϵ.*
- *if $p_s \in [U_s,\infty)$: $n^*(s,p_s,\epsilon) = 0$.*

We give the proof of Proposition 1 in appendix [24]. Proposition 1 reveals that users' optimal prompt strategy exhibits non-monotonic variation with ambiguity ϵ. The behavior can be categorized into three distinct cases, depicted in Figure 4:

- **Extreme Cases:** When $p_s = 0$ or $p_s \ge U_s$, users either continuously prompt the model $s\in\mathcal{M}$ or completely avoid AIGC services of model $s\in\mathcal{M}$.

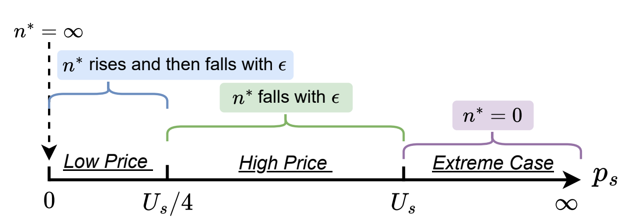


Fig. 4. The Optimal Prompt Number $n^*(s,p_s,\epsilon)$ of Users with Varying Levels of Prompt Ambiguity $\epsilon\in(0,1)$.

- **Low Price Range** ($p_s \in (0, U_s/4]$): The optimal prompt number follows an inverse U-shaped pattern as ambiguity rises. Users with moderate prompt capability will use more prompts than those with very high or low capabilities, due to higher marginal utility of additional prompts.
- **High Price Range** ($p_s \in (U_s/4, U_s)$): Decreased prompt capability (higher ambiguity ϵ) consistently reduces the optimal prompt number. The elevated price amplifies the disadvantage of lower prompt capability, discouraging users with higher prompt ambiguity from AIGC services.

C. The User's Optimal Expected Payoff

Given the prompt strategy $n^*(s, p_s, \epsilon) \in \mathbb{N}$ in Theorem 1 and GAI model decision $s^*(n^*, \boldsymbol{p}) \in \mathcal{M}$ in (8), we summarize the user's optimal payoff $\pi^*_{\text{user}}(s^*, n^*, \boldsymbol{p}, \epsilon)$ as follows:

$$\pi^*_{\text{user}}(s^*, n^*, \boldsymbol{p}, \epsilon) = \left(1 - \epsilon^{n^*(s^*, p_{s^*}, \epsilon)}\right) \cdot U_{s^*} - n^*(s^*, p_{s^*}, \epsilon) \cdot p_{s^*}, \tag{10}$$

where U_{s^*} and p_{s^*} are the utility mapping and prompt price of the user's optimal model $s^*(n^*, \boldsymbol{p}) \in \mathcal{M}$, respectively.

With $\pi^*_{\text{user}}(s^*, n^*, \boldsymbol{p}, \epsilon)$ in (10), we aim to discuss the effect of prompt ambiguity $\epsilon \in (0, 1)$ on the user's optimal expected payoff, as concluded in Proposition 2.

Proposition 2. *The user's optimal payoff $\pi^*_{\text{user}}(s^*, n^*, \boldsymbol{p}, \epsilon)$ decreases with prompt ambiguity $\epsilon \in (0, 1)$.*

Given the complex nature of users' two-step decisions (as illustrated in Proposition 1), the proof of Proposition 2 requires careful analysis and is provided in detail in appendix [24]. This proposition indicates a crucial insight: while users may adapt their prompt strategy $n^*(s, p_s, \epsilon)$ in various ways depending on the ratio p_s/U_s (particularly when $p_s/U_s \in (0, 1/4]$), their optimal payoff $\pi^*_{\text{user}}(s^*, n^*, \boldsymbol{p}, \epsilon)$ invariably decreases as ambiguity ϵ increases. This demonstrates that despite users' strategic adjustments to maximize their payoff under different ambiguity levels, the negative impact of increased prompt ambiguity on the optimal payoff cannot be fully mitigated.

Having completed our analysis of heterogeneous users' optimal decisions and payoffs, we now proceed to examine the optimal prompt pricing in Stage 1 of AIGC services.

IV. Platform's Prompt Pricing Mechanism Design

This section derives the platform's optimal prompt pricing mechanism by examining two distinct user scenarios. First, we focus on users with consistent prompt ambiguity, i.e., homogeneous cases, to establish a closed-form solution for optimal pricing. Then, based on these theoretical results, we extend our analysis to heterogeneous users and develop an efficient prompt pricing algorithm for platform payoff maximization.

A. The Optimal Prompt Pricing for Homogeneous Users

For a published GAI model set $\mathcal{M}$, the platform aims to determine the optimal prompt pricing $\boldsymbol{p}^* = \{p^*_m, \forall m \in \mathcal{M}\}$ that maximizes its payoff $\pi_{\text{platform}}(\boldsymbol{p})$ in (6), formulated as a bi-level optimization in Problem 1. Theorem 2 establishes the optimal prompt price for each model when serving homogeneous users with prompt ambiguity $\epsilon \in (0, 1)$.

Theorem 2. *For a model set $\mathcal{M}$ where each model $m \in \mathcal{M}$ has utility mapping U_m and prompt cost C_m, the platform's optimal prompt pricing $\boldsymbol{p}^* = \{p^*_m, \forall m \in \mathcal{M}\}$ for homogeneous users with prompt ambiguity $\epsilon \in (0, 1)$ is:*

$$p^*_m = \begin{cases} \epsilon^{\sigma(m,\epsilon)-1}(1-\epsilon)U_m, & \text{if } m = m^*, \quad (11\text{a}) \\ U_m, & \text{if } m \neq m^*, \quad (11\text{b}) \end{cases}$$

where $\sigma(m, \epsilon) = \min\{k \in \mathbb{N} : (k+1)\epsilon^k - k\epsilon^{k-1} < \frac{C_m}{(1-\epsilon)U_m}\}$ and $m^* = \operatorname{argmax}_{m \in \mathcal{M}}(\epsilon^{\sigma(m,\epsilon)-1}(1-\epsilon)U_m - C_m) \cdot \sigma(m, \epsilon)$.

We give the proof of Theorem 2 in the online appendix [24]. Theorem 2 formalizes the optimal prompt price p^*_m of each model $m \in \mathcal{M}$, which increases with utility mapping U_m and decreases with prompt cost C_m.

Building on Theorem 2, and in preparation for designing prompt pricing with heterogeneous users, we further need to understand the impact of ambiguity ϵ on optimal pricing $\boldsymbol{p}^*$.

Corollary 1. *For model $m^* \in \mathcal{M}$ with utility mapping U_{m^*}, prompt cost C_{m^*}, and optimal prompt price $p^*_{m^*}$, the user's optimal strategy $n^*(m^*, p^*_{m^*}, \epsilon)$ varies with prompt ambiguity $\epsilon \in (0, 1)$ as follows:*

- *if $C_{m^*} = 0$: $n^*(m^*, p^*_{m^*}, \epsilon)$ increases with ϵ.*
- *if $C_{m^*} \in (0, U_{m^*}/8]$: $n^*(m^*, p^*_{m^*}, \epsilon)$ first increases and then decreases with ϵ.*
- *if $C_{m^*} \in (U_{m^*}/8, U_{m^*})$: $n^*(m^*, p^*_{m^*}, \epsilon)$ decreases with ϵ.*
- *if $C_{m^*} \in [U_{m^*}, \infty)$: $n^*(m^*, p^*_{m^*}, \epsilon) = 0$.*

We give the proof of Corollary 1 in appendix [24]. This relationship between prompt cost and optimal strategy is illustrated in Figure 5. In zero-cost scenario ($C_{m^*} = 0$), increasing ambiguity leads the platform to encourage more prompts to maximize payoff. For low cost ($C_{m^*} \in (0, U_{m^*}/8]$), the platform initially reduces prices as ambiguity grows to stimulate higher prompt numbers, but beyond a threshold, it raises prices, reducing prompt usage. With higher cost ($C_{m^*} \in (U_{m^*}/8, U_{m^*})$), increased ambiguity consistently leads to fewer prompts under optimal pricing, as incentivizing additional prompts becomes unprofitable.

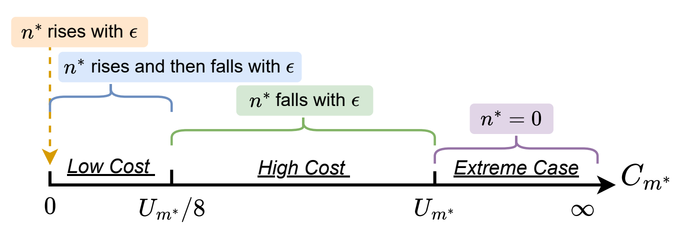


Fig. 5. The Optimal Prompt Number $n^*(m^*, p_{m^*}, \epsilon)$ of Homogeneous Users with Different Ambiguity ϵ under $p^*_{m^*}$ in (11a).

Through Corollary 1, Proposition 3 highlights the impact of prompt ambiguity $\epsilon \in (0, 1)$ on the platform's optimal payoff, corresponding to that on users in Proposition 2.

Proposition 3. *The platform's optimal payoff $\pi^*_{platform}(\boldsymbol{p}^*)$ under homogeneous users decreases with ambiguity $\epsilon \in (0, 1)$.*

The complete proof of Proposition 3 is provided in appendix [24]. This proposition implies that prompt ambiguity has a detrimental effect on both the platform's and users' payoffs. Notably, an increase in prompt ambiguity ϵ does not guarantee a higher number of user prompts, as shown in Proposition 1. Even though the platform strategically optimizes the balance between prompt pricing and user prompt frequency, both parties experience diminishing optimal payoffs, creating a suboptimal outcome for all stakeholders.

B. The Optimal Prompt Pricing for Heterogeneous Users

Having analyzed the homogeneous case, we now address the more practical scenario of AIGC services with heterogeneous user prompt ambiguity. We begin by conducting a theoretical analysis of Problem 1 within this heterogeneous context, which is a multi-variable non-convex optimization (strongly NP-hard even in linear forms [29]). Then, we exploit the interaction between user behavior, GAI models, and prompt pricing to design an efficient algorithm for solving it.

1) Theoretical Analysis: Based on heterogeneous users' optimal prompt strategy in Theorem 1, we first investigate the payoff maximization for the platform under any single model $m \in \mathcal{M}$. With Proposition 1 and Lemma 2, we reformulate such a problem into an unconstrained piecewise optimization in Problem 2, where $\lambda_1(k) \in (0, 1)$ and $\lambda_2(k) \in (0, 1)$ are the implicit solutions of equation $\epsilon^{k-1}(1-\epsilon)U_m = p_m$ with $\lambda_1(k) \leq \lambda_2(k)$, and upper bound $\bar{n}^*(m)$ in (9).

Problem 2 (Prompt Pricing for Single GAI Model)**.**

$$\max_{p_m} (p_m - C_m)\Big(\int_0^{1-\frac{p_m}{U_m}} f(\epsilon)d\epsilon + \sum_{k=2}^{\bar{n}^*} \int_{\lambda_1(k)}^{\lambda_2(k)} f(\epsilon)d\epsilon.$$

This transformation significantly reduces the complexity of optimal prompt pricing for single model. We illustrate this with Example 1, using a uniform distribution of prompt ambiguity.

Example 1 (Uniform Distribution of User Prompt Ambiguity)**.** *For prompt ambiguity uniformly distributed over $[\epsilon_{min}, \epsilon_{max}]$ where $\epsilon_{min} < \epsilon_{max}$:*

- *Problem 2 exhibits strict concavity on each piecewise segment $k \in [2, \bar{n}^*(m)]$.*
- *When $\epsilon_{min} = 0$ and $C_m > 0$, the optimal prompt price is $p^*_m = (U_m + C_m)/2$.*

Example 1 reveals that under uniform ambiguity distribution, the platform prioritizes users with lower ambiguity to maximize $\pi_{\text{platform}}(\boldsymbol{p})$. When the ambiguity distribution's lower bound is zero, the optimal price p^*_m depends solely on utility mapping U_m and prompt cost C_m.

Following the prompt pricing for single GAI model, we extend our exploration to the entire model set $\mathcal{M}$ and incorporate users' optimal model decisions derived in (8). Specifically, for any two models m and m' in $\mathcal{M}$, we establish upper bound $\bar{p}_m(m', \epsilon)$ on prompt price p_m such that users with prompt ambiguity $\epsilon \in (0, 1)$ will prefer model $m \in \mathcal{M}$ over $m' \in \mathcal{M}$, and vice versa, expressed as below:

$$\bar{p}_m(m', \epsilon) = \Big[(1 - \epsilon^{\tau(m,m',\epsilon)}) \cdot U_m + n^*(m', p_{m'}, \epsilon) \cdot p_{m'} - (1 - \epsilon^{n^*(m', p_{m'}, \epsilon)}) \cdot U_{m'}\Big] \cdot \frac{1}{\tau(m, m', \epsilon)}, \tag{12}$$

where $n^*(m', p_{m'}, \epsilon)$ is in Theorem 1 and $\tau(m, m', \epsilon)$ is:

$$\tau(m, m', \epsilon) = \min\{k \in \mathbb{N}^+ : (1 - \epsilon^{n^*}) \cdot U_{m'} - n^* \cdot p_{m'} < (1 - \epsilon^k) \cdot U_m - k\epsilon^k(1 - \epsilon) \cdot U_m\}. \tag{13}$$

From (12), the optimal prompt pricing involves multi-level interactions with heterogeneous users' prompt strategy and model decision, especially given model set $\mathcal{M}$ and ambiguity distribution $f(\epsilon)$. This interrelation introduces non-convexity and significant computational complexity, making it difficult to derive a closed-form solution in general scenarios. Hence, we next aim to develop an efficient algorithm for Problem 1.

2) Algorithm Design: To address Problem 1, we develop the Optimal Prompt Pricing (OPP) algorithm (Algorithm 1) through systematic decomposition into tractable sub-problems. Our approach, inspired by finite element methods [30]–[32], enhances computational efficiency by reducing the search space to a single dimension compared to exhaustive search. The algorithm integrates price upper bounds and strategically partitions the ambiguity distribution based on user decisions.

Workflow of Algorithm 1: Given prompt cost and utility mapping of GAI models in set $\mathcal{M} = \{M_L, M_H\}$, and user prompt ambiguity density function $f(\epsilon)$, we identify the upper bound $\bar{p}_H(M_L, \epsilon)$ in (12), corresponding to prompt price p_L of model M_L (Line 2-3). Then, we partition $f(\epsilon)$ according to the optimal model decisions of heterogeneous users in AIGC services, formulated as a function of prompt price p_H:

$$\tilde{f}(\epsilon, p_H) = \begin{cases} f(\epsilon), & \text{if } p_H \leq \bar{p}_H(M_L, \epsilon), \quad (14a) \\ 0, & \text{if } p_H > \bar{p}_H(M_L, \epsilon). \quad (14b) \end{cases}$$

Algorithm 1: The Optimal Prompt Pricing (OPP).

Input: prompt cost C_L and C_H, utility mapping U_L and U_H, user prompt ambiguity density function $f(\epsilon)$, and step size α.

Output: the optimal prompt pricing $\boldsymbol{p}^* = \{p^*_L, p^*_H\}$ and resultant payoff π^*_{platform}.

1 $\pi^*_{\text{platform}} = 0$;
2 **for** $p_L = C_L : \alpha : U_L$ **do**
3   Derive $\bar{p}_H(M_L, \epsilon)$ in (12);
4   Construct $\tilde{f}(\epsilon, p_H)$ in (14) and $G(p_H)$ in (15);
5   $\hat{p}_H = \{x \in [C_H, \bar{p}_H(M_L, 0)] : \partial G(x)/\partial x = 0\}$;
6   $p_H = \text{argmax}_{x \in \{C_H, \hat{p}_H, \bar{p}_H(M_L, 0)\}} G(x)$;
7   **if** $\pi_{platform}(p_L, p_H) \geq \pi^*_{platform}$ **then**
8     $\boldsymbol{p}^* = \{p_L, p_H\}$;
9     $\pi^*_{\text{platform}} = \pi_{\text{platform}}(p_L, p_H)$;
10 Return $\boldsymbol{p}^*$ and π^*_{platform};

Using revised ambiguity density function $\tilde{f}(\epsilon, p_H)$ from (14), we construct $G(p_H)$ in (15), aligning it with optimizing price p_H for platform payoff maximization (Line 4):

$$G(p_H) = \int_0^1 \tilde{f}(\epsilon, p_H) \cdot [(p_H - C_H) \cdot n^*(M_H, p_H, \epsilon) \\ -(p_L - C_L) \cdot n^*(M_L, p_L, \epsilon)] \, d\epsilon. \quad (15)$$

Through the preceding transformation and simplification, we derive the optimal prompt price p_H^* for any given p_L of model M_L (Line 5-6). By iteratively updating the prompt pricing mechanism, the OPP algorithm ultimately outputs the optimal one and resultant payoff for the platform (Lines 7-9).

V. Experimental Evaluation

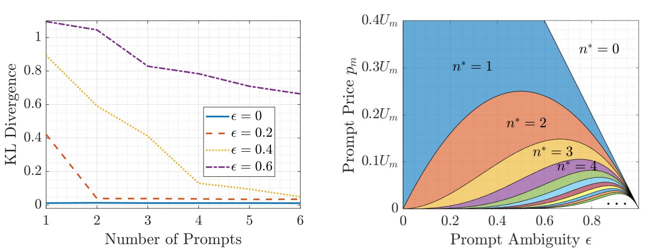
This section outlines the experimental setup and presents results validating the theoretical analysis of user strategies and prompt pricing, highlighting their impact on platform payoff.

A. Experimental Setup

We adopt a character-level GPT-like model [23], with 21 million parameters and a block size of 256 to simulate practical AIGC services. Following [20]–[22], we generate synthetic language data for training using a doubly-embedded Markov chain model. To simulate varying user prompt ambiguity ϵ, we inject noise into the transition matrix and uniformly distribute prompt ambiguity within $[\epsilon_{\text{min}}, \epsilon_{\text{max}}]$. We scale system factors proportionally to GAI models' utility to analyze their interactions. Due to the page limit, we include additional results, such as those for different ambiguity distributions, in appendix [24], where the insights are similar.

B. The User's Optimal Strategy and Expected Payoff

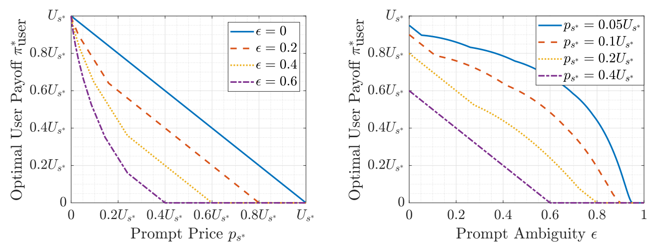
1) The User's Optimal Strategy: Figure 6 illustrates how prompt ambiguity ϵ and the number of prompts n influence the GAI model's output and the user's optimal strategy. In Figure 6(a), the KL divergence [33] between the model's output and the user's true intention increases with ambiguity ϵ but decreases as the number of prompts n grows. When $\epsilon = 0$ (no ambiguity), the divergence is minimal, aligning with the analysis in Section II-B. Figure 6(b) shows that users with moderate ambiguity tend to issue more prompts, while a higher prompt price p_m reduces the number of prompts.



(a) Prompt Number and Ambiguity (b) Optimal Number $n^*(m, p_m, \epsilon)$

Fig. 6. The User's Optimal Utilization Strategy.

2) The User's Optimal Payoff: Figure 7 demonstrates how the user's optimal payoff π^*_{user}, given the model's optimal decision s^*, varies with ambiguity ϵ and prompt price p_{s^*}. As seen in Figure 7(a), ambiguity ϵ determines the price threshold p_{s^*} at which the user achieves a positive payoff and is willing to purchase the AIGC service. Figure 7(b) reveals that as p_{s^*} increases, the user's payoff π^*_{user} declines despite strategic adjustments in utilization. Notably, higher ambiguity ϵ leads to a sharper reduction in the user's optimal payoff.

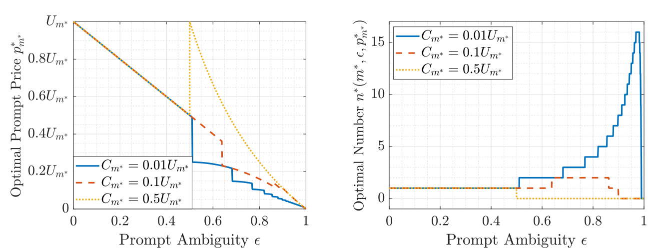


(a) Payoff π^*_{user} and Price p_{s^*} (b) Payoff π^*_{user} and Ambiguity ϵ

Fig. 7. The User's Optimal Expected Payoff.

C. The Platform's Optimal Pricing and Expected Payoff

1) The Platform's Optimal Pricing: Figure 8 examines the platform's optimal prompt price $p^*_{m^*}$ for model m^* under varying prompt ambiguities ϵ and the resulting user strategy $n^*(m^*, p^*_{m^*}, \epsilon)$. In Figure 8(a), the optimal pricing behavior depends on prompt cost C_{m^*} and ambiguity ϵ. Initially, as ϵ increases, the platform lowers $p^*_{m^*}$ to encourage more prompts and maximize its payoff. However, beyond a critical ambiguity threshold (e.g., $\epsilon = 0.86$ for $C_{m^*} = 0.1U_{m^*}$, marked by the red dashed line in both subfigures), the platform shifts to increasing $p^*_{m^*}$, prioritizing higher prices over more frequent prompts for payoff maximization.



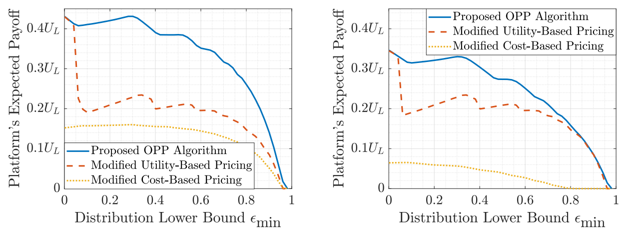
(a) Price $p^*_{m^*}$ and Ambiguity ϵ (b) $n^*(m^*, p^*_{m^*}, \epsilon)$ with $p^*_{m^*}$

Fig. 8. The Platform's Optimal Prompt Pricing.

2) The Platform's Optimal Payoff: Figure 9 compares the platform's payoff under the OPP algorithm for model set $\mathcal{M} = \{M_L, M_H\}$ with two benchmarks:

- Modified Utility-Based Pricing [34]: Adapts value-based pricing but neglects users' strategic decisions across model sets.
- Modified Cost-Based Pricing [35]: Incorporates prompt costs but overlooks heterogeneous user strategies, limiting payoff optimization.

As shown in Figure 9, the platform's payoff decreases with lower ϵ_{min} under all pricing mechanisms. While both OPP and utility-based pricing exhibit fluctuations due to user strategies, the OPP algorithm consistently outperforms benchmarks by leveraging users' two-step decisions. It achieves at least 75.05% and 31.72% higher payoffs in the setups of Figures 9(a) and 9(b), respectively.



(a) Setup of $U_H = 1.8U_L$, $C_H = 0.04U_L$ and $C_L = 0.02U_L$

(b) Setup of $U_H = 1.5U_L$, $C_H = 0.06U_L$ and $C_L = 0.02U_L$

Fig. 9. The Platform's Optimal Expected Payoff ($\epsilon_{\max} = 1$).

VI. Conclusion

This paper provides the first study on strategic interactions between users and the platform in AIGC services. By introducing prompt ambiguity, we derive the optimal user utilization strategies and platform prompt pricing. Our exploration reveals complex interactions among decisions and system factors, where ambiguity adversely affects the payoff of both parties.

In future works, we will refine our theoretical analysis and OPP algorithm to support GAI models with more granular performance tiers, improving adaptability to diverse applications. Meanwhile, we will enhance our experiments with practical data sets and model-based agents.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Project 62472367, 62102343, 62271434), Guangdong Research (Project 2021QN02X778), Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (No. ZDSYS20220606100601002), Shenzhen Stability Science Program 2023, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects (No. LGKCSDPT2024002), Shenzhen Science and Technology Program (Project JCYJ20230807114300001, JCYJ20220818103006012), and Suzhou Frontier Science and Technology Program (Project SYG202310).

References

[1] A. Karapantelakis *et al.*, "Generative AI in mobile networks: a survey," *Annals of Telecommunications*, vol. 79, pp. 15–33, 2024.
[2] "Monica - your ChatGPT AI assistant for anywhere," https://monica.im/, accessed: 2025-1-24.
[3] "Poe - fast, helpful AI chat," https://poe.com, accessed: 2025-1-24.
[4] OpenAI *et al.*, "Gpt-4 technical report," 2024.
[5] OpenAI. (2023) Gpt-4 turbo (gpt-4o). Accessed: 2025-1-24. [Online]. Available: https://openai.com
[6] Y. Cao *et al.*, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," 2023. [Online]. Available: https://arxiv.org/abs/2303.04226
[7] H. Du *et al.*, "Enabling ai-generated content services in wireless edge networks," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 226–234, 2024.
[8] P. Dütting *et al.*, "Mechanism design for large language models," in *ACM Web Conference*, 2024, pp. 144–155.
[9] M. Xu *et al.*, "Sparks of gpts in edge intelligence for metaverse: Caching and inference for mobile aigc services," 2023. [Online]. Available: https://arxiv.org/abs/2304.08782
[10] Y. Liu *et al.*, "Prosecutor: Protecting mobile aigc services on two-layer blockchain via reputation and contract theoretic approaches," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 10 966–10 983, 2024.
[11] H. Du *et al.*, "Exploring collaborative distributed diffusion-based ai-generated content (aigc) in wireless networks," *IEEE Network*, vol. 38, no. 3, pp. 178–186, 2024.
[12] G. Liu *et al.*, "Semantic communications for artificial intelligence generated content (aigc) toward effective content creation," *IEEE Network*, vol. 38, no. 5, pp. 295–303, 2024.
[13] J. Wu *et al.*, "Ai-generated content (aigc): A survey," 2023. [Online]. Available: https://arxiv.org/abs/2304.06632
[14] C. Zhang *et al.*, "A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?" 2023. [Online]. Available: https://arxiv.org/abs/2303.11717
[15] Y. Wang *et al.*, "A survey on chatgpt: Ai–generated contents, challenges, and solutions," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 280–302, 2023.
[16] "Bearly," https://bearly.ai, accessed: 2025-1-24.
[17] P. Sahoo *et al.*, "A systematic survey of prompt engineering in large language models: Techniques and applications," 2024. [Online]. Available: https://arxiv.org/abs/2402.07927
[18] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, 2022, pp. 24 824–24 837.
[19] Q. Dong *et al.*, "A survey on in-context learning," in *Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 1107–1128.
[20] Y. Zhou *et al.*, "The mystery of in-context learning: A comprehensive survey on interpretation and analysis," in *Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 14 365–14 378.
[21] S. M. Xie *et al.*, "An explanation of in-context learning as implicit bayesian inference," 2022. [Online]. Available: https://arxiv.org/abs/2111.02080
[22] H. Jiang, "A latent space theory for emergent abilities in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.09960
[23] A. Radford *et al.*, "Improving language understanding by generative pre-training," *OpenAI*, 2018.
[24] X. Li *et al.*, "Online appendix: Strategic prompt pricing for aigc services: A user-centric approach," https://drive.google.com/drive/folders/1JFq3SoqNZwMC8I-yPEEHybfBOupp3OtJ?usp=sharing, accessed: 2025-1-24.
[25] D. Patel and A. Ahmad, "The inference cost of search disruption – large language model cost analysis," https://semianalysis.com/2023/02/09/the-inference-cost-of-search-disruption/, accessed: 2025-1-15.
[26] R. Pieraccini and E. Levin, "Stochastic representation of semantic structure for speech understanding," *Speech Communication*, vol. 11, no. 2, pp. 283–288, 1992.
[27] S. Miller *et al.*, "Hidden understanding models of natural language," in *Annual Meeting on Association for Computational Linguistics*, 1994, pp. 25–32.
[28] A. Mas-Colell *et al.*, *Microeconomic Theory*. Oxford University Press, 1995.
[29] P. Hansen *et al.*, "New branch-and-bound rules for linear bilevel programming," *SIAM Journal on Scientific and Statistical Computing*, vol. 13, no. 5, pp. 1194–1217, 1992.
[30] W. Bangerth and R. Rannacher, *Adaptive finite element methods for differential equations*. Springer Science & Business Media, 2003.
[31] B. Szabó and I. Babuška, *Finite element analysis: Method, verification and validation*. John Wiley & Sons, 2021.
[32] I. Babuška and T. Strouboulis, *The finite element method and its reliability*. Oxford university press, 2001.
[33] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
[34] J. Pei, "A survey on data pricing: From economics to data science," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4586–4608, 2022.
[35] C. Courcoubetis and R. Weber, *Pricing communication networks: Economics, technology and modelling*. John Wiley & Sons, 2007.