Evaluation of Agents under Simulated AI Marketplace Dynamics

To Eun Kim
Carnegie Mellon University
Pittsburgh, PA, United States
toeunk@cs.cmu.edu

Alireza Salemi
University of Massachusetts Amherst
Amherst, MA, United States
asalemi@cs.umass.edu

Hamed Zamani
University of Massachusetts Amherst
Amherst, MA, United States
zamani@cs.umass.edu

Fernando Diaz
Carnegie Mellon University
Pittsburgh, PA, United States
diazf@acm.org

Abstract

Modern information access ecosystems consist of mixtures of systems, such as retrieval systems and large language models, and increasingly rely on marketplaces to mediate access to models, tools, and data, making competition between systems inherent to deployment. In such settings, outcomes are shaped not only by benchmark quality but also by competitive pressure, including user switching, routing decisions, and operational constraints. Yet evaluation is still largely conducted on static benchmarks with accuracy-focused measures that assume systems operate in isolation. This mismatch makes it difficult to predict post-deployment success and obscures competitive effects such as early-adoption advantages and market dominance. We introduce *Marketplace Evaluation*, a simulation-based paradigm that evaluates information access systems as participants in a competitive marketplace. By simulating repeated interactions and evolving user and agent preferences, the framework enables longitudinal evaluation and marketplace-level metrics, such as retention and market share, that complement and can extend beyond traditional accuracy-based metrics. We formalize the framework and outline a research agenda, motivated by business and economics, around marketplace simulation, metrics, optimization, and adoption in evaluation campaigns like TREC.

CCS Concepts

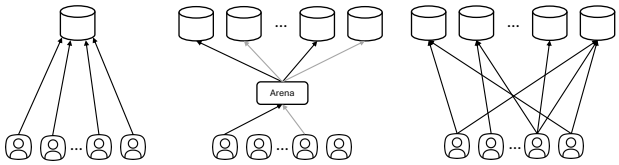
• **Computing methodologies** → **Modeling and simulation**; *Natural language generation*; • **Information systems** → *Information retrieval*.

Keywords

Simulated Evaluation, Information Access Agents, Marketplace Dynamics

ACM Reference Format:
To Eun Kim, Alireza Salemi, Hamed Zamani, and Fernando Diaz. 2026. Evaluation of Agents under Simulated AI Marketplace Dynamics. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3808542

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808542



(a) Cranfield **(b) Arena** **(c) Marketplace**

Figure 1: Comparison of evaluation paradigms for information access systems. (a) Cranfield: a multi-user single-system setting where all queries are routed to a fixed system, eliminating competition and user choice. (b) Arena: a multi-user multi-system setting where a central platform mediates anonymous pairwise comparisons, precluding persistent user preferences or switching behavior. (c) Marketplace: a multi-user multi-system setting where users actively select among competing providers without requiring a central mediator, enabling competitive dynamics, behavioral adaptation, and the emergence of market share over time.

1 Introduction

Modern information access (IA) systems are increasingly interactive, often operating through conversational interfaces powered by large language models (LLMs) that deliver generated information objects (GIOs) to users [20]. These systems, hereafter referred to as *IA agents*, can be any system that provides information access functionality, including standalone LLMs, standalone retrieval systems, or compositions of one or more LLMs and retrievers [97].

Development in this space has been rapid and highly competitive, with providers introducing new capabilities at a near-monthly pace [56, 62]. This competition plays out across marketplaces that mediate access to various information access agents, including open community hubs [62], centralized execution platforms [7, 69], and enterprise marketplaces [6, 24]. The existence of these platforms make it easy for users and systems to try multiple providers and switch when better options appear.

However, evaluation of IA agents is predominantly conducted using static benchmarks that treat systems as standalone artifacts [74, 90]. This isolated view is therefore misaligned with deployment: within marketplace, performance is inherently interdependent, since the success of one system depends not only on its intrinsic quality but also on the presence, behavior, and strategic choices of competing systems. These interactions fundamentally alter what

it means for a system to be *good*, shifting evaluation from absolute effectiveness to relative performance under competitive pressure.

Understanding system performance in the competitive marketplaces is crucial for both system designers and platform operators, as isolated benchmarks often fail to predict real-world adoption and success. For example, in retrieval-augmented generation (RAG) [58], the interaction between generators and retrievers creates network effects where the value of a retrieval system depends not only on its intrinsic quality but also on its adoption by the generator population and its differentiation from competing systems [73]. Marketplace dynamics introduce phenomena such as winner-take-all effects [79], where small performance differences can lead to dramatic market share disparities, and portfolio effects [64], where generators may strategically diversify across multiple retrieval systems to optimize for different query types or hedge against system failures. These dynamics have significant implications for research priorities, business strategies, and the overall health of information ecosystems, as they influence innovation incentives, market concentration, and the diversity of available approaches.

We present the perspective that existing evaluation paradigms in information access, including Cranfield-style evaluations [18] and pairwise A/B testing [45], do not account for competitive market forces or user behavioral adaptation. In the Cranfield paradigm (Figure 1a), evaluation can be interpreted as a multi-user single-system setting, where multiple queries stand in for multiple users but all interactions are routed to a fixed system, eliminating any notion of competition or user choice. Pairwise and Arena-style evaluations (Figure 1b) extend this to a multi-user multi-system setting, yet still fall short of modeling real-world competition, since users are typically presented with anonymous system variants and do not actively select among providers, preventing the formation of persistent preferences or switching behavior. As a result, the single-system evaluation fails to capture how system performance degrades or improves under varying load conditions as market share fluctuates, while pairwise comparisons cannot model the complex multi-way interactions that emerge when multiple systems compete simultaneously for the same user base. Furthermore, current approaches do not account for strategic behavior by system operators, such as quality differentiation that influence system selection beyond pure performance metrics. The temporal dynamics of marketplace competition, including learning effects, adaptation strategies, and the evolution of user preferences, are entirely absent from traditional evaluation frameworks, yet these factors critically determine long-term system viability and market outcomes.

Following this perspective, we introduce *marketplace evaluation*, a general framework for assessing information access systems as participants in competitive marketplaces rather than as isolated artifacts. The framework models information access as a multi-agent system in which providers and intermediaries make strategic decisions under shared constraints, and system performance emerges from interactions among competing entities rather than from standalone effectiveness.

Without loss of generality, we ground this framework in the RAG ecosystem, a representative setting that encompasses all key interactive stakeholders, including users, generators, and retrievers. In this setting, for interactive information access, users choose among multiple competing generators, which differ in capability and cost, while generators in turn act as demand-side agents that select among competing retrieval services. Here, retrieval systems also compete for traffic coming from generators not only through retrieval quality, but also through capability, cost, and control over which collections of information they are able or permitted to serve. Both generators and retrievers may strategically restrict, expand, or specialize their accessible information sources, making access itself a central axis of competition.

Within this instantiation, we formalize evaluation with an agent-based simulation [10, 27, 76] of competitive information access marketplaces, in which system performance depends on adoption patterns, competitive pressure, and information access constraints. Our framework incorporates realistic marketplace factors such as capacity limits, pricing mechanisms, and quality-of-service guarantees, enabling systematic evaluation of how systems perform, adapt, and coexist under different competitive structures.

We begin in Section 2 with our perspective and a motivating experiment that illustrates how competitive dynamics alter system-level conclusions. We then formalize the marketplace evaluation framework with consistent notations in Section 3. Building on this foundation, we outline a long term research program centered on three core research questions:

- RQ1 (§4): How can information access agent marketplaces be simulated for evaluation?
- RQ2 (§5): What metrics are appropriate for characterizing IA agent's performance and market conditions?
- RQ3 (§6): How can marketplace evaluation be integrated into existing evaluation campaigns and benchmarking infrastructures?

We hope this fresh perspective inspires a new class of methodologies for evaluating and improving the impact of information retrieval technologies on AI ecosystems, while addressing one of the most pressing challenges in modern AI.

2 Perspective: Evaluation of IA Agents as Marketplace Participants

We argue that IA agents should be evaluated as competitors in a marketplace, where performance emerges through repeated interaction and both human and agent preferences evolve over time. In such environments, systems compete for attention, usage, and continued engagement from users. Consequently, evaluation outcomes depend on how systems are exposed to users, how preferences form over time, and how competition shapes long-term usage patterns.

Current evaluation methods are fundamentally limited by the assumption that interactions are independent and isolated, failing to capture how humans and AI agents evolve through experience. Traditional static benchmarks and Cranfield paradigm ignore the fact that human preferences are shaped by prior exposure. In the real world, the specific sequence and timing of interactions build the loyalty or abandonment that dictates market success [96]. Similarly, modern multi-agent IA systems often utilize *routers* [42, 55, 68, 85] to dynamically direct queries to the most appropriate retriever or generator. These components also develop their own internal tendencies over time by learning from feedback such as cost, latency, and accuracy. Consequently, they exhibit *path dependence* where early interactions alter future system behavior and resource allocation. By treating every interaction as a first time occurrence, current

paradigms miss the cumulative, historical dynamics that define real-world performance. This misalignment motivates a rethinking of evaluation paradigms toward settings and metrics that explicitly model and evaluate interaction, adaptation, and competition.

This evaluation perspective algins with common business and deployment goals. For industry practitioners, the success of an IA agents is rarely defined by a single offline metric [32, 41]. Instead, deployment success is reflected through operational signals such as query traffic, market share, and sustained usage [78]. However, this perspective does not diminish the importance of accuracy-based metrics (e.g., NDCG [89] for ranking evaluation, and ROUGE [59] for GIO evaluation), as those quality do influence user satisfaction and underlies downstream outcomes such as retention and growth [22]. Marketplace-level metrics therefore complement the traditional metrics, rather than replacing them.

2.1 Motivating Experiment

To illustrate what evaluation under competition reveals that traditional static benchmarks cannot and to motivate a future research agenda, we evaluate multiple IA agents using a marketplace simulation with a simple setup.

We first conduct static evaluations to create a ranking of systems. We set up seven distinct systems: DeepSeek V3.2 [60], Kimi K2.5 [86], Gemini 2.5 [19], GPT-OSS [1], Grok 4.1 [91], Qwen3 [93], and Llama 3.3 [26]. These agents are tested using 500 fact-seeking questions drawn from the test set of the SimpleQA benchmark. [90]. These agents produce generated information objects (GIO) [20] instead of being restricted to short-phrase responses. Each agent answers all questions, responses are scored using a question-level correctness metric, and systems are ranked by the aggregated metric (F-score), following the benchmark. This procedure yields a single, static system ranking that is invariant to evaluation order and independent of user interaction. The evaluation results and the system ranking can be found in Table 1.

2.1.1 ***Marketplace Simulation Setup.*** We contrast this static evaluation with a minimal marketplace simulation that preserves the same questions and the same correctness metric, but introduces user choice and preference updating. We simulate a population of 10 users who can access all systems in the market and maintain individual preference distributions over systems. Per simulation step, 5 users are sampled, and each user is assigned to a question drawn without replacement from the dataset pool, and the user selects a system based on their current preference distribution with a small amount of exploration. After observing the system response, the user updates their preference based solely on the question-level correctness of that response.

This way, the first 100 simulation steps cover the same 500 questions used in our static evaluation. We initially run these steps with six models, introducing a seventh model during the second half of the simulation (steps 101 to 200). This setup enables us to examine two contrasting scenarios under an active market with warm-started users. In the first, a system that performs strongly under static benchmarking, Qwen3, enters a lightly concentrated market (Table 2 & Figure 2a). In the second, a system with moderate static benchmark performance, DeepSeek V3.2, enters a highly concentrated market (Table 3 & Figure 2b).

Table 1: Model performance (F1-score) and fair market shares proportionate to the benchmark performance.

Model	F1	Fair Share (%)	Fair Share (%) w/o Qwen3	Fair Share (%) w/o DeepSeek V 3.2
Qwen 3	60.24	28.89	-	33.30
Kimi K2.5	42.51	20.38	28.66	23.49
Llama 3.3	27.74	13.30	18.71	15.33
DeepSeek V3.2	27.59	13.23	18.60	-
Grok 4.1	19.21	9.21	12.95	10.62
Gemini 2.5	16.83	8.07	11.35	9.30
GPT-OSS	14.42	6.91	9.72	7.97

2.1.2 ***Ranking Divergence Under Interaction.*** Let us begin by comparing the system rankings derived from the static benchmark (Table 1) with those obtained from the first set of simulation runs prior to the entry of any new models ($t = 1$–100) in Tables 2 and 3. Although the same set of questions was used, the resulting rankings differ, indicating that competitive dynamics and user adaptation reshape outcomes beyond what is captured by static evaluation alone. This discrepancy serves as one indicator that marketplace interactions can meaningfully alter system standing even before any structural changes, such as new entry, occur.

This becomes even clearer in examining the market share trends in Figures 2a and 2b. Market share is defined as the total query traffic received by an IA agent within a given time window. The plots report windowed market share with a size of 10, allowing us to track local trends as time progresses. When systems are ranked by cumulative market share over $t = 1$–100, a single overall ordering emerges. However, the windowed curves reveal that local rankings fluctuate frequently. In several periods, the market share of a model drops close to zero before recovering, or vice versa. These fluctuations highlight the underlying dynamics of user adaptation and competition that are not visible in aggregate statistics alone.

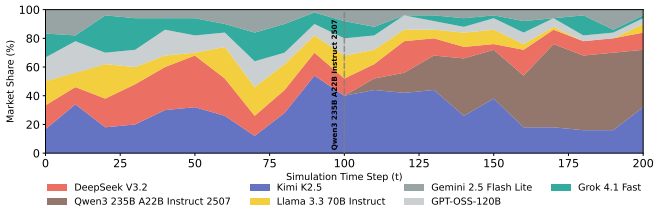
2.1.3 ***Market Entry and Dominance Effects.*** More interesting effects emerge when the marketplace participants changes. From the windowed market share plots at $t = 100$, the market in Figure 2a is visibly less concentrated than in Figure 2b. The entry of Qwen3 acts as a clear shock to the market, whereas the entry of DeepSeek V3.2 produces only modest shifts, suggesting that entering an already highly concentrated market makes it difficult for a new model to secure the share implied by its standalone performance.

From a meritocratic perspective, one might expect market share to be proportional to internal capability, as approximated by static benchmark performance. Table 1 reports the models' "fair" shares derived from their F1 scores. However, when we compare these expected fair shares (FS) with the realized market shares, the discrepancy is substantial (ΔFS in Tables 2 and 3). In particular, post entry market share disparity is notably higher than pre entry disparity, as reflected by the HHI index reported in Tables 2 and 3.

In Figure 2a, the entry of Qwen3 compresses the market shares of middle and lower ranked models into nearly indistinguishable levels. At that stage, the ranking implied by static benchmarking becomes largely uninformative. Across both scenarios, we observe dominance effects that resemble real world marketplaces and cannot be inferred from static benchmarking alone. The marketplace metrics discussed here, including market share and HHI, are examined in greater detail in Section 5.

Table 2: Qwen3 late entry.

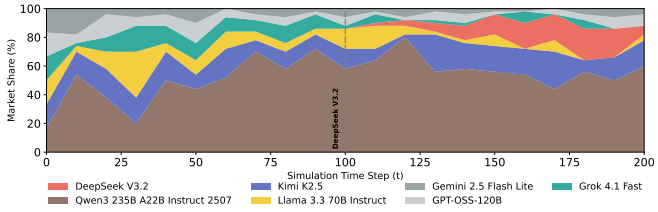
Market Share (ΔFS) (%)	System Ranking	→	System Ranking	Market Share (%) (ΔFS)
Before Entry HHI = 1940.96 $t = (1-100)$			After Entry HHI = 2430.16 $t = (101-200)$	
		new	Qwen3	36.00 (+7.11)
(+0.74) 29.40	Kimi K2.5		Kimi K2.5	29.40 (+8.62)
(+2.4) 21.00	DeepSeek V3.2		DeepSeek V3.2	11.60 (-1.63)
(-4.31) 14.40	Llama 3.3		Gemini 2.5	6.60 (-1.47)
(+0.85) 13.80	Grok 4.1		GPT-OSS	6.40 (-0.51)
(+3.08) 12.80	GPT-OSS		Llama 3.3	5.60 (-7.7)
(-2.75) 8.60	Gemini 2.5		Grok 4.1	4.40 (-4.81)



(a) Qwen3-235B-A22B-Instruct-2507 introduced at $t = 100$. Late entry of a strong model into a lightly concentrated market.

Table 3: DeepSeek V3.2 late entry.

Market Share (ΔFS) (%)	System Ranking	→	System Ranking	Market Share (%) (ΔFS)
Before Entry HHI = 3176.88 $t = (1-100)$			After Entry HHI = 3789.28 $t = (101-200)$	
(+22.71) 51.60	Qwen3		Qwen3	57.80 (+28.91)
(-5.58) 14.80	Kimi K2.5		Kimi K2.5	15.80 (-4.58)
		new	DeepSeek V3.2	12.0 (-1.23)
(-2.70) 10.60	Llama 3.3		Llama 3.3	4.60 (-8.70)
(+0.39) 9.60	Grok 4.1		GPT-OSS	4.20 (-2.71)
(+0.49) 7.40	GPT-OSS		Gemini 2.5	3.20 (-4.87)
(-2.07) 6.00	Gemini 2.5		Grok 4.1	2.40 (-6.81)



(b) DeepSeek V3.2 introduced at $t = 100$. Late entry of an average model into a highly concentrated market.

Figure 2: Marketplace dynamics illustrating fluctuating system rankings and market share following the entry of a new model. The first 100 simulation steps were conducted with six models, after which a seventh model was introduced into the warm started marketplace.

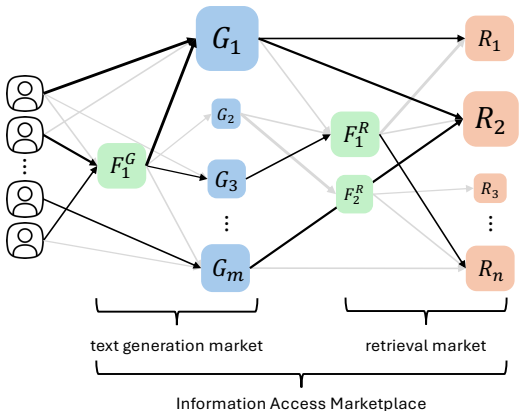


Figure 3: An example snapshot of a RAG marketplace simulation. Nodes represent agents grouped by stakeholder role: users, generators, retrievers, and routers to generators and retrievers. Arrow thickness reflects selection preference intensity, node size reflects accumulated market share, and arrow opacity indicates whether a connection was selected or not selected in the current snapshot.

3 The Marketplace Evaluation Framework

Before we delve into the core research questions, in this section, we formalize the marketplace evaluation framework, including how interaction dynamics unfold across successive simulation steps.

3.1 Overview

To study and evaluate systems within complex ecosystems, we formalize our framework using agent-based simulation [10, 27, 76]. The dynamics observed in competitive information access agent marketplaces arise from interactions among many decision-makers, including human users and agents, coupled through feedback, competition, and path dependence. For such systems, analytical characterization quickly becomes intractable and/or insoluble. As argued by Axtell [10] and, in the context of business research by Rand and Rust [76], agent-based simulation is not merely a modeling convenience but a necessary methodology for studying complex systems in which aggregate behavior emerges from interactions among heterogeneous agents.

Moreover, simulation provides a controlled yet expressive foundation for evaluation [11, 92]. It enables the study of both short- and long-term deployment effects, supports scalable and reproducible experimentation without reliance on live user studies, and allows situated and counterfactual analysis under specific user populations and competitive conditions [39, 94]. Recent advances in large language models make such simulations increasingly feasible [31, 52, 70, 99], positioning simulation-based evaluation as a practical mean to measure agents in the real-world marketplaces.

Figure 3 illustrates a concrete example of the agent-based simulation setting, involving users, generators, retrievers, and routers. Consider a domain expert and a general web user interacting with three generator options: a high-capability general model, a moderate general-purpose model, and a domain-specialized model. The generators can access two retrieval back-ends: a public web search retriever and a proprietary domain retriever such as a legal document index. A router mediates requests from generators to retrievers based on routing policies. Generators may either route queries through the router or directly call a retriever, creating heterogeneous interaction patterns that evolve over time.

In this RAG ecosystem simulation example [97], a single interaction unfolds as follows: (i) a user issues a request and either directly

selects, or is routed to, a generator; (ii) the selected generator may in turn select a retriever to obtain supporting evidence before producing a response; (iii) the resulting output is evaluated to produce a utility signal, which is then used to update future selection behavior of both users and generators. As this process repeats across many interactions and many users, traffic allocation, agent preferences, and market share evolve over time.

Marketplace evaluation focuses on these *longitudinal outcomes* with an ecosystem-centric view [92]. Rather than scoring agents in isolation on one-shot effectiveness, we evaluate how agents attract demand, retain usage, and remain viable under competition. The remainder of this section formalizes the components and dynamics underlying such simulations.

Before turning to our core research agenda, we first formalize the agent-based marketplace simulation used for dynamic evaluation.

3.2 Marketplace Primitives

Stakeholders. We define a *stakeholder* as a population of agents sharing a common *functional role* in the marketplace. Grounding this definition in the example from Section 3.1, the users, generators, retrievers, and router each constitute distinct stakeholder populations within the same marketplace. Roles here are defined by the task an agent performs, rather than by its internal architecture. For example, the same language model may act as a generator when producing responses, or as a retriever when ranking documents. Let $\mathcal{A} = \{\mathcal{U}, \mathcal{G}, \mathcal{R}, \mathcal{F}, \ldots\}$ denote the universe of possible stakeholder populations (users, generators, retrievers, routers, etc.). A concrete evaluation instantiates a finite subset $A \subseteq \mathcal{A}$. Each stakeholder $s \in A$ contains a set of agents with internal state parameters $\{\theta_{s,a}\}_{a \in s}$. We describe common stakeholders and representative algorithmic choices from the literature in Section 4.

Markets and the Information Access Marketplace. Competition arises whenever multiple agents provide functionally substitutable services for the same task and therefore compete for traffic. In the example from Section 3.1, competition for user traffic arises not only among generators and among retrievers, but also across roles, as both generators and retrievers seek to capture and retain demand within the marketplace. We define a *market* as a task-specific group of stakeholders that jointly realize a functionality and compete for exposure within that functionality. An *information access marketplace* consists of one or more such markets. Prominent examples include a *text generation market* (generators competing to respond to user queries) and a *document retrieval market* (retrievers competing to supply evidence), as depicted in Figure 3.

Between different markets, collaboration can arise through compositional workflows. Retrieval-enhanced generation [97] illustrates how the retrieval market and the text generation market collaborate to improve end-user outcomes by incorporating retrieved evidence into generation. At the same time, competition *between* markets can still emerge when users choose among different information access modalities (e.g., selecting a conversational search versus a web search).

Marketplace Governance Graph. Inspired by Zhuge et al. [101] and Hoveyda et al. [36], the interaction structure of an instantiated marketplace (Figure 3) is defined by a directed acyclic graph

$$G_{\text{marketplace}} = (A, \mathcal{E}), \tag{1}$$

where nodes correspond to instantiated stakeholders in A. An edge $(x \to y) \in \mathcal{E}$ indicates an *admissible selection (invocation)* relation: an agent in stakeholder x may select (i.e., call, route to, or delegate to) an agent in stakeholder y as part of its computation. Equivalently, the governance graph can be represented by an adjacency matrix

$$W \in \{0, 1\}^{|A| \times |A|}, \tag{2}$$

where $W[x, y] = 1$ iff $(x \to y) \in \mathcal{E}$, otherwise $W[x, y] = 0$. Importantly, $G_{\text{marketplace}}$ is a *marketplace governance graph* specifying which selections are allowed at the level of stakeholder roles, rather than the (possibly iterative) execution trace within a single episode.

Interaction Semantics and Dynamics. We now formalize how interactions unfold within an instantiated marketplace. Given a fixed marketplace governance graph and stakeholder populations, interaction semantics specify how agents are selected, how artifacts are produced, and how realized outcomes induce feedback and adaptation. This formulation abstracts over architectural details and captures execution at the level of *who interacts with whom*, rather than *how* each agent internally computes its output.

Each stakeholder population $s \in A$ is associated with a stochastic policy $\Pr_s(\cdot \mid \text{pa}(s); \theta_s)$, where $\text{pa}(s) = \{s' \in A \mid (s' \to s) \in \mathcal{E}\}$ denotes the set of parent stakeholders of s in $G_{\text{marketplace}}$, and θ_s denotes the parameters of agents within s. A single interaction instance is generated by sampling along the governance graph in topological order:

$$z \sim \prod_{s \in \text{topo}(G_{\text{marketplace}})} \Pr_s(\cdot \mid \text{pa}(s); \theta_s), \tag{3}$$

where z denotes the realized trajectory of selected agents.

Given an interaction outcome z, evaluation produces a utility:

$$\mu \sim \Pr(\cdot \mid z, \mathcal{D}; \theta_\mu), \tag{4}$$

where $\mathcal{D}$ denotes the evaluation dataset and θ_μ denotes the evaluator model. Depending on the dataset, evaluation may be reference-based (e.g., relevance or accuracy [90]) or reference-free (e.g., LLM-based judges [28]), depending on the application.

After observing μ, participants may update their internal states:

$$\theta_{s^*} \leftarrow \textsc{Update}(s^*, z, \mu), \tag{5}$$

where s^* denotes any agent selected during z. Referring to the example in Section 3.1, suppose the expert user selects the moderate general-purpose generator, which in turn relies on the general web retriever. If the returned result is unsatisfactory, the user may reduce its preference for this generator in expert-level tasks. This negative feedback can propagate upstream, discouraging the generator from selecting the same retriever in similar contexts. Conversely, positive outcomes can reinforce the current routing and selection behavior across all involved agents. Repeated interaction, evaluation, and adaptation induces a stochastic dynamical system over the instantiated stakeholders and agents.

Table 4: Prominent stakeholders in information access agent marketplaces, defined by functional roles. Each stakeholder adapts its behavior based on feedback from the simulated marketplace, and evaluation is conducted using longitudinal and market-level signals that complement traditional effectiveness metrics.

Stakeholder	Goal	Adaptive Behavior in the Marketplace	Evaluation Focus
Users	Sustained utility from the marketplace over time	Adapt preferences over generators based on past satisfaction, cost, latency, or trust	Longitudinal utility, retention rate, switching behavior
Generators	Maximize user demand under quality–cost tradeoffs	Adapt interaction protocols and retrieval configurations to remain competitive	User utility, cost efficiency, market share, user retention
Routers	Efficient allocation of queries to downstream services	Exclusive routing (vertical integration) vs. open routing (market-wide discovery)	Counterfactual regret, allocation efficiency, diversity, and fairness
Retrievers	Remain selected by generators under competition	Compete via domain specialization, personalization, or robustness	Marginal utility contribution, selection frequency, long-term survival

4 RQ1: Marketplace Simulation

To understand how IA agent marketplace should be simulated, we question how we can effectively model users (RQ 1.1) and agents (RQ 1.2). To do so, we discuss each stakeholder and their objectives in the marketplace. Table 4 summarizes the primary stakeholders in an IA marketplace. Rather than viewing these stakeholders as static system components, we model them as *strategic actors* that pursue objectives, adapt their behavior under competition, and are evaluated based on their long-run impact on marketplace outcomes.

Across all stakeholder types, we need to consider:

(1) a goal that defines what success means for the stakeholder,
(2) adaptation strategies that determine how behavior changes over time, and
(3) evaluation signals that reflect whether those strategies succeed in the marketplace.

4.1 RQ1.1: Modeling Users

4.1.1 ***Users: Demand-Side Adaptation.*** Users represent the demand side of the marketplace. Their objective is to maximize utility from interactions, accounting for factors such as answer quality [35], cost [2], and trust [48]. Referring to the example in Section 3.1, the expert user may be willing to incur higher cost to access a more capable generator, whereas the general user may prioritize lower cost or faster responses. As a result, each user type operates under a different notion of utility.

Based on their notion of utility, users may adapt by reallocating their demand across competing generators based on past experience [43, 63]. This adaptation may take the form of gradual preference shifts, exploration of alternatives after poor outcomes, or abandonment of previously favored services.

Accordingly, users can be evaluated indirectly, through the signals they generate [35]. Longitudinal utility trajectories, retention rates, and switching behavior can serve as market-level feedback that shapes competition among downstream stakeholders.

4.1.2 ***Simulation of Users.*** Grounded in the stakeholder analysis above, user behavior in the simulation can be modeled through an explicit utility function. For a user u interacting with agent a at time t, utility may be defined as $\theta_{\mu}^{u,a}(t) = \alpha_u\, Q_{u,a}(t) - \beta_u\, C_a(t) - \gamma_u\, L_a(t)$, where Q denotes answer quality, C cost, L latency, and the coefficients capture heterogeneous user sensitivities and encode tradeoff across competing attributes.

Given realized utility, user preferences over agents can be viewed in a perspective of choice modeling across heterogeneous agents [88]. Different agents can pursue distinct objectives under shared constraints, and their interactions jointly determine marketplace outcomes. In choice modeling, decisions are formalized as utility maximization over bundles of commodities [33], directly aligning with our formulation. One simple option, is stochastic choice model [88] that assigns selection probabilities to agents as a smooth function of their estimated utility.

Preference modeling choices are tightly coupled to the construction of the dataset $\mathcal{D}$. While $\mathcal{D}$ may be instantiated using existing benchmarks or synthetically generated queries [15, 63, 75, 98], the structure and topical distribution of the dataset implicitly determines the space over which user preferences are defined. This raises another question: whether user preference should be expressed by topic-dependent distribution, or at the level of persistent users whose preferences evolve across simulation steps.

Prior work has proposed taxonomies of IA behaviors [16, 30, 83] which can be derived from interaction logs [46, 100]. However, it remains unclear how these should be operationalized as sampling distributions or state variables within θ_u when users face multiple substitutable services. Understanding this coupling between dataset design and preference dynamics is central to modeling user adaptation in a marketplace with multiple substitutable services.

Another interesting direction concerns cross market competition formed by users. Users may substitute across IA stakeholders based on task characteristics. For example, they may choose a traditional search engine over than a conversational system for certain queries [14]. Competition therefore occurs not only within a market, but also across markets with partially substitutable functionality.

Beyond structural choices in user simulation, marketplace evaluation invites a deeper examination of behavioral assumptions. Behavioral economics [65] suggests that, within the user population, repeated exposure to multiple substitutable agents may introduce effects that are central to behavioral economics, such as framing [66], constructed preferences [88], and network effects [87]. Incorporating these phenomena into θ_u raises questions about which behavioral mechanisms materially affect marketplace outcomes. For example, popularity signals or exposure biases [96] introduced by their selection policies may amplify market concentration.

4.2 RQ1.2: Modeling Agents

4.2.1 ***Generators: Competing for Demand under Constraints.*** Generators are often supply-side actors interfacing directly with users, especially for GIO creations [20]. From business perspective [41], their primary objective is to attract and retain users while operating under quality-cost tradeoffs [50] imposed by inference scaling, retrieval, or orchestration choices [23]. Looking at the example in Section 3.1, a more capable generator may prioritize answer quality over traffic, as its higher per-query cost can still yield substantial revenue per user, whereas moderate models may rely on higher traffic volume to compensate for lower revenue per interaction.

Such generators can be modeled as optimizing an expected utility function that balances user-derived utility signals with operational cost. Adaptation then corresponds to updating retrieval strategies, orchestration depth, or model configuration to maximize long-run utility or market share.

Within the marketplace, generators may operate without retrieval [9], rely on a single retriever [58], aggregate evidence from multiple retrievers [47], or engage in more complex multi-step or agentic retrieval strategies [17, 44].

Evaluation of generators should therefore extend beyond answer quality to include sustained user utility, market share [62, 69], cost efficiency [51], and retention.

4.2.2 ***Routers: Market Intermediaries.*** Routers mediate access to downstream services by allocating queries among competing agents. In the Section 3.1's example, if the generator lacks an effective updating policy, it may depend on the router for retriever selection. However, if routing decisions consistently yield poor outcomes, the generator may bypass it in subsequent interactions.

The two common routers: user-to-generator routers [3, 37, 38, 68] and generator-to-retriever routers [47, 55, 85, 95] differ in position but share a common objective.

A router can be modeled as optimizing an allocation utility that aggregates downstream performance, cost, and fairness constraints. Preference updates correspond to adjusting routing probabilities based on observed utility or regret signals over time.

Routers may adopt exclusive or open routing policies [69], shaping exposure and competitive pressure. Evaluation should therefore focus on allocation-level outcomes such as assignment efficiency [42], regret [13, 85], and exposure diversity or fairness [25, 54].

4.2.3 ***Retrievers: Competing for Downstream Selection.*** Retrievers supply evidence to generators and compete for repeated selection. Their objective can be modeled as maximizing marginal contribution [81] to generator utility while maintaining efficiency and robustness.

Retriever adaptation may manifest as domain specialization [12]. Because retrievers are tied to underlying collections, differences in coverage can implicitly correspond to different data providers. Retrievers may also differentiate through personalization to generators [82]. In this setup, retrievers that fail to provide consistent marginal value [81] will gradually lose traffic.

Evaluation can therefore focus on marginal contribution, as well as robustness [61] and efficiency [8, 57].

5 RQ2: Marketplace Metrics

In Section 2, together with the experiment (§2.1), we highlighted the limitations of traditional accuracy-based metrics for evaluating agents deployed in competitive environments. While relevance, correctness, and utility remain foundational, they do not capture how agents perform *as market participants*, where success is reflected in sustained usage, traffic allocation, and competitive positioning. This motivates the need for *marketplace metrics* that characterize both individual agent outcomes and emergent marketplace dynamics.

Under RQ2, we examine how to evaluate performance in a marketplace setting. Specifically, we distinguish between metrics that measure the market performance of individual agents (RQ 2.1) and metrics that characterize overall market conditions from the perspective of platform operators (RQ 2.2).

Accordingly, we propose two complementary classes of metrics. *Agent-Level Metrics* quantify the position and performance of individual agents within the marketplace, such as their share of traffic and their ability to retain users. *Marketplace-Level Metrics* capture aggregate structural properties, including concentration and inequality, which reflect dominance, competitiveness, and overall market health.

5.1 RQ2.1: Agent-Level Metrics

As shown in Figure 2, we derive agent market share from the interaction logs of the simulation. Rather than computing market share cumulatively over the entire horizon, we use a sliding window to capture temporal dynamics and short-term competitive effects.

Let A denote the population of information access agents in a given market, and let $q_u(t) \in A$ denote the agent selected by user $u \in U$ at timestep t. For a window of length w, the market share (MS) of agent $a \in A$ at time t is defined as

$$\mathrm{MS}_a(t; w) = \frac{\sum_{u \in U} \sum_{k=t-w+1}^{t} \mathbf{1}[q_u(k) = a]}{\sum_{u \in U} \sum_{k=t-w+1}^{t} 1}, \tag{6}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, which equals 1 when its argument is true and 0 otherwise. Therefore, $\mathrm{MS}_a(t; w)$ measures the fraction of total interaction volume allocated to agent a within the window. Tracking $\mathrm{MS}_a(t; w)$ over time reveals how agents gain or lose traffic as users adapt their preferences and routing policies evolve.

Beyond traffic volume, an agent's long-term viability depends on its ability to retain users after initial adoption. Inspired by customer retention metrics from the business literature [53], we define customer retention (CR) measures tailored to agent marketplaces.

Let $\tau_{u,a} = \min\{\, t \in \{1, \ldots, T\} : q_u(t) = a \,\}$ denote the first timestep at which user u selects agent a, if such a timestep exists. Given a window length m, the *user-level retention* of agent a for user u is

$$\mathrm{CR}_{u,a}(m) = \frac{1}{m} \sum_{t=1}^{m} \mathbf{1}\left[q_u(\tau_{u,a} + t) = a\right], \tag{7}$$

which measures the fraction of the next m interactions, following first adoption, in which u continues selecting a. Aggregating across users yields the *agent-level retention*

$$\mathrm{CR}_a(m) = \frac{1}{|U_a|} \sum_{u \in U_a} \mathrm{CR}_{u,a}(m), \tag{8}$$

where $U_a := \{u \in U : \tau_{u,a} \text{ is defined}\}$ is the set of users who have tried agent a at least once. A higher $\mathrm{CR}_a(m)$ indicates that users who initially sample a tend to continue allocating a larger fraction of their subsequent interactions to that agent.

However, higher market share does not necessarily imply higher customer retention. A model may maintain a small market share yet achieve strong retention by consistently delivering high quality to a targeted user segment, as in Section 3.1, where a more capable model provides exclusive value to the domain expert.

5.2 RQ2.2: Marketplace-Level Metrics

Beyond agent-level metrics, marketplace-level metrics characterize aggregate interaction patterns and diagnose higher-level market properties such as concentration, convergence, and inequality. These metrics can be computed from the cumulative interaction behavior observed in the simulation logs. These marketplace-level metrics primarily reflect the interests of mediation platform owners or routers rather than individual agents. If the market becomes highly concentrated, routing may become trivial or even unnecessary, and competing service providers may progressively lose their share of demand. Here, standard market concentration and inequality metrics from the economics literature can be applied [62].

The Herfindahl–Hirschman Index (HHI) [77] captures the degree of market concentration. Using the windowed market-share from Equation (6), we define HHI-based market concentration as

$$\mathrm{HHI}(t; w) = \sum_{a \in A} \mathrm{MS}_a(t; w)^2. \tag{9}$$

Larger $\mathrm{HHI}(t; w)$ indicates greater concentration, approaching monopoly as the value increase.

Marketplace outcomes may also be evaluated through fairness lenses, e.g., whether providers receive exposure consistent with a target allocation. Notion of *fair* market share helps us creating related marketplace-level metrics. Let's define a *target* share or exposure distribution ϵ^* over agents, where $\sum_{a \in A} \epsilon^*_a = 1$. Depending on the definition of fairness, we can define ϵ^*_a as $1/|A|$ for equal share, or proportional to the merit of an agent, where the merit can be based on the static benchmark score (as in Table 1) or cumulated μ's during the window w at time t.

From here, we can yield a complementary market dominance measure

$$\Delta(t; w) = \max_{a \in A} \mathrm{MS}_a(t; w) - \epsilon^*_a. \tag{10}$$

A larger $\Delta(t; w)$ indicates that the top agent captures substantially more market share than the target share.

Additionally, inspired by expected exposure (EE) in document ranking [25], we can define an expected exposure difference metric by comparing realized traffic shares to a target exposure distribution ϵ^* over agents. Using market share as a proxy for exposure, we define expected exposure difference as

$$\mathrm{EE}_{\text{marketplace}}(t; w) = \|\mathrm{MS}_a(t; w) - \epsilon^*_a\|_2^2, \tag{11}$$

where $\|\cdot\|_2^2$ denotes the squared ℓ_2 norm.

Following the definition of expected exposure disparity (EE-D) [25], we obtain $\text{EE-D}_{\text{marketplace}}(t; w) = \|\mathrm{MS}_a(t; w)\|_2^2$, which is analogous to the $\mathrm{HHI}(t; w)$.

6 RQ3: Adoption to Evaluation Campaigns

The marketplace evaluation framework also raises questions about how large-scale evaluation campaigns, such as TREC [84] and CLEF [29], can evolve beyond traditional Cranfield-style setups. Historically, these campaigns have relied on a static test collection and offline run submission, where systems are evaluated independently by organizers or assessors. While this paradigm has enabled reproducibility and straightforward comparability, it does not capture the dynamics we want to measure.

Adopting a simulated marketplace environment within evaluation campaigns introduces new design choices that can be organized along two orthogonal axes.

Axis 1: Peer Competition Vs. Benchmark Competition. In a peer competition setting, submitted systems compete directly with one another for traffic and exposure within a shared marketplace simulation. Outcomes will be jointly determined by the participant pool, enabling analysis of dominance, switching, and concentration under head to head conditions. However, results may depend on the specific mix of systems submitted in a given year, potentially limiting stability and cross-year comparability. In a benchmark competition setting, each submission competes against a fixed population of organizer-provided IA agents. This improves reproducibility, allowing clearer attribution of outcomes to individual system properties.

Axis 2: Run Submission Vs. Agent Submission. In the run submission model, participants provide static outputs that are treated as cached agent responses, and only the simulated user population adapts over time. This preserves many practical advantages of traditional campaigns, including low computational burden, and reproducibility. However, systems themselves cannot adapt or respond to competitive feedback. In the agent submission model, participants submit executable agents that interact with a shared simulation, making sequential decisions under competition. This enables evaluation of adaptive routing, cost-aware strategies, and strategic behavior, but introduces challenges in controlling stochasticity, ensuring fair resource allocation, and maintaining reproducibility.

Task Design. From a task-design perspective, marketplace-based campaigns could focus on settings where interaction and adaptation are essential, such as distributed information retrieval (e.g., TREC Million LLMs [49]), generator–retriever coordination (e.g., TREC RAG [74]), and interactive information access (e.g., TREC iKAT [4]). These tasks would allow campaigns to evaluate not only effectiveness, but also robustness, adaptability, and strategic behavior under competition.

7 Validation of Marketplace Simulation

Validation of marketplace simulations is a necessary step following any simulated evaluation [5, 76]. Because simulation outcomes emerge from interacting components rather than isolated system runs, validation must assess whether the constructed environment meaningfully reflects the phenomena it aims to study.

We organize validation around three pillars. Validation of (1) sampled or synthetically generated user queries; (2) marketplace interaction dynamics and stakeholder modeling; and (3) marketplace metrics (meta-evaluation).

Validation of User Queries. Validation of user queries concerns whether sampled or synthetically generated queries adequately represent the intended demand distribution. From a construct validity perspective, the query distribution should align with the real user population it claims to model. This can be examined through distributional comparisons against real logs when available [46, 83], or through expert review of coverage across task types, difficulty levels, and user expertise strata.

Convergent validity can be assessed by measuring system rank correlations between simulation-based evaluations and established offline benchmarks. Suppose we obtain systems from a public model leaderboard that reports download counts or market share. We can instantiate those systems within our marketplace simulation and run the simulation under controlled conditions. Based on a chosen marketplace metric, we can produce a ranking of systems from the simulation. We then compute the correlation between the simulated ranking and the ranking implied by the real-world leaderboard. A higher correlation would suggest that the simulation captures relevant aspects of real-world competitive dynamics.

Validation of Marketplace Stakeholders. As another example of construct validity, we can introduce controlled perturbations within the marketplace setup and examine whether the intended marketplace-level or agent-level metrics respond accordingly. For instance, if we artificially increase the latency of a particular model, we would expect to observe lower customer retention or increased market concentration due to the system's degradation.

Meta Evaluation of Marketplace Metrics. Marketplace metrics require meta evaluation to ensure they meaningfully support comparison [80]. Key properties include sensitivity, discrimination power, and robustness to gaming. Sensitivity can be assessed through bootstrap based hypothesis testing and ASL curves. Discrimination power measures a metric's ability to separate agents with genuinely different performance profiles (e.g., if the current definition of customer retention rate has a strong discrimination power is unclear). Robustness to gaming evaluates whether agents can inflate scores without improving underlying utility. Grounded in measurement theory, these analyses strengthen the credibility of marketplace simulation as an evaluation framework.

8 Further Research Agenda

8.1 Single-Market Vs. Multi-Market

As shown in Section §2.1, evaluation can be conducted using only user-facing agents without explicitly modeling downstream interactions. Recall a *market* denotes a task-specific set of stakeholders that jointly deliver a functionality and compete for traffic within it. We refer to such settings as *single-market* evaluations. Here, routers that mediate user-to-generator selection remain within the same market. Although front-facing agents may rely on complex internal components, evaluation considers only their observable end outcomes (e.g., final responses or rankings), abstracting away internal or downstream dynamics.

Single-market evaluation can be naturally extended to *multi-market* evaluation by explicitly modeling multiple coupled markets within the same simulation. As exemplified by RAG, such settings involve distinct markets that are linked through a compositional workflow (Figure 3).

In contrast to single-market evaluation, multi-market evaluation explicitly models inter-market coupling and captures how competition, mediation, and composition jointly shape user-facing outcomes. For example, a generator's realized quality depends on the retriever it invokes, while a retriever's value depends on how generators use its evidence. By treating routers and retrievers as first-class stakeholders and logging full compositional trajectories, including which agents were selected in each market and with what outcomes, multi-market evaluation supports cross-market diagnosis and enables attribution of observed behaviors to specific markets or mediation policies.

8.2 Economic Models of Competition

Broadly, the marketplace perspective motivates closer engagement with economic models of competition [88]. Classical concepts such as market dominance [34] and competitive entry [72] have well-established theoretical foundations, but their application to AI and information access agent marketplaces remains largely unexplored. Also, as discussed in Section 4.1.2, incorporating choice modeling for agent selection and behaviorally grounded user and agent models [65] remains underexplored in information retrieval. Such perspectives can enrich the study of retrieval in *multifirm* settings.

Beyond being helpful to information retrieval research, this framework may also benefit scholars in computational economics and algorithmic market design. By providing a controllable simulation environment with explicit user adaptation, agent competition, and market-level observables such as market share, retention, and concentration, it offers a testbed for studying dynamic competition among AI agents under realistic demand feedback. Such a setting enables the empirical exploration of classical economic constructs, including dominance, entry, switching costs, and equilibrium formation, in domains where analytical modeling alone may be insufficient. In this sense, marketplace simulation and evaluation framework provides a new computational laboratory for economic research on digital competition.

8.3 Optimization of Agents and Adversaries

Marketplace evaluation naturally raises the question of how information access agents should be optimized when success is defined by competitive, market-level outcomes along with effectiveness metrics. Simulated environments have played an important role in optimizing agents in interactive settings [71], particularly in reinforcement learning, where feedback is delayed and behavior unfolds over time.

In a marketplace, optimization can be framed as an online decision problem where agents repeatedly choose actions (e.g., agent selection [49]) under uncertainty about competitor behavior. This setting naturally supports the study of bandit [85] and online learning-to-rank methods [21, 102] for agent selection that optimize long-term objectives such as retention, or profit under cost constraints.

The optimization of agents in a marketplace naturally brings attention to adversarial behaviors. Agents may deliberately or unintentionally exploit weaknesses in evaluation metrics, routing policies, or user models, leading to outcomes that are harmful to the marketplace as a whole. This includes behaviors analogous to reward hacking [40], where agents optimize for observed signals

while undermining the intended objectives of the system. Understanding such behaviors raises questions about the robustness of marketplace metrics, the vulnerability of routing mechanisms, and the safeguards required to ensure healthy competition [67].

8.4 Implementation Choices

Marketplace simulations require several practical design decisions that can affect outcomes. Key choices include the number of users, and the number of competing systems. The batch size of users to sample per simulation step may influences the effective time scale of adaptation. The simulation horizon determines whether results reflect short term dynamics or long term convergence. Finally, from sampled users, simulations may operate synchronously, with global in-batch updates, or asynchronously, with staggered interactions that more closely resemble real deployment.

8.4.1 Resource Contribution. To facilitate systematic exploration of these factors, in the interest of feasibility and reproducibility, we release a Python package `marketplace-eval` that implements the proposed framework. The package provides a minimal starting point with example configurations, including the exact simulation setup used in our motivating experiment, enabling researchers to readily replicate results and extend to new settings. Beyond parameter control over population sizes, batch sampling, horizon length, and interaction mode, the package supports modular overrides of core components, allowing researchers to customize interaction dynamics and adapt the framework to a wide range of experimental settings. [1]

9 Conclusion

We propose a marketplace-based perspective for evaluating information access agents, shifting evaluation from isolated system performance to outcomes shaped by interaction, adaptation, and competition. We outline a research agenda and release extensible resources to support further study of marketplace evaluation. This perspective encourages evaluation frameworks that better reflect real-world deployment, where information access systems co-evolve with users and other agents.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, and in part by the National Science Foundation under Grant No. 2402873 and 2402874. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

[1] The code is available at https://github.com/kimdanny/marketplace-eval.

References

[1] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925* (2025).

[2] Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. 2024. AutoMix: automatically mixing language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 4164, 35 pages.

[3] Shubham Agrawal and Prasang Gupta. 2025. LLMRank: Understanding LLM Strengths for Model Routing. *arXiv preprint arXiv:2510.01234* (2025).

[4] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. TREC iKAT 2023: A Test Collection for Evaluating Conversational and Interactive Knowledge Assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 819–829. https://doi.org/10.1145/3626772.3657860

[5] Mary J. Allen and Wendy M. Yen. 2001. *Introduction to Measurement Theory*. Waveland Press, Long Grove, IL.

[6] Amazon Web Services. 2026. *AWS Marketplace*. https://aws.amazon.com/marketplace/solutions/ai-agents-and-tools Accessed: 2026-02-04.

[7] Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol Accessed: 2026-02-05.

[8] Nima Asadi and Jimmy Lin. 2013. Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '13)*. Association for Computing Machinery, New York, NY, USA, 997–1000. https://doi.org/10.1145/2484028.2484132

[9] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[10] Robert Axtell. 2000. *Why agents?: on the varied motivations for agent computing in the social sciences*. Vol. 17. Center on Social and Economic Dynamics Washington, DC.

[11] Krisztian Balog and ChengXiang Zhai. 2024. User Simulation for Evaluating Information Access Systems. *Foundations and Trends® in Information Retrieval* 18, 1-2 (2024), 1–261. https://doi.org/10.1561/1500000098

[12] Ryan C Barron, Vesselin Grantcharov, Selma Wanna, Maksim E Eren, Manish Bhattarai, Nicholas Solovyev, George Tompkins, Charles Nicholas, Kim Ø Rasmussen, Cynthia Matuszek, et al. 2024. Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization. In *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1669–1676.

[13] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[14] Kevin Matthe Caramancion. 2024. Large language models vs. search engines: evaluating user preferences across varied information retrieval scenarios. *arXiv preprint arXiv:2401.05761* (2024).

[15] Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. 2011. Simulating Simple User Behavior for System Effectiveness Evaluation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. Association for Computing Machinery, New York, NY, USA, 611–620.

[16] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt*. Technical Report. National Bureau of Economic Research.

[17] Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. BeamAggR: Beam Aggregation Reasoning over Multi-source Knowledge for Multi-hop Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1229–1248. https://doi.org/10.18653/v1/2024.acl-long.67

[18] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. *Aslib Proceedings* 19, 6 (2025/06/06 1967), 173–194.

[19] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[20] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (Aug. 2018), 34–90. https://doi.org/10.1145/3274784.3274788

[21] Zhuyun Dai, Yubin Kim, and Jamie Callan. 2017. Learning to rank resources. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*. 837–840.

[22] Ovidiu Dan and Brian D Davison. 2016. Measuring and predicting search engine users' satisfaction. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 1–35.

[23] Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. 2025. Multi-Agent Collaboration via Evolving Orchestration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=L0xZPXT3le

[24] Databricks, Inc. 2026. *Databricks Marketplace*. https://www.databricks.com/product/marketplace Accessed: 2026-02-04.

[25] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[26] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[27] Joshua M Epstein and Robert Axtell. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

[28] Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 150–158.

[29] Nicola Ferro and Carol Peters (Eds.). 2019. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. The Information Retrieval Series, Vol. 41. Springer Cham. https://doi.org/10.1007/978-3-030-22948-1

[30] Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Q&A Benchmarks for RAG Evaluation in Enterprise Settings, Georg Rehm and Yunyao Li (Eds.). Association for Computational Linguistics, Vienna, Austria, 469–484. https://doi.org/10.18653/v1/2025.acl-industry.33

[31] Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?. In *Proceedings of the 25th ACM Conference on Economics and Computation*. ACM, New Haven CT USA, 614–615. https://doi.org/10.1145/3670865.3673513

[32] Carlos A. Gomez-Uribe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (Dec. 2016), 19 pages. https://doi.org/10.1145/2843948

[33] William Greene. 2009. *Discrete Choice Modeling*. Palgrave Macmillan UK, London, 473–556. https://doi.org/10.1057/9780230244405_11

[34] Joseph E. Harrington and Myong-Hun Chang. 2005. Co-evolution of firms and consumers and the implications for market dominance. *Journal of Economic Dynamics and Control* 29, 1-2 (Jan. 2005), 245–276. https://doi.org/10.1016/j.jedc.2003.04.012

[35] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*. 221–230.

[36] Mohanna Hoveyda, Harrie Oosterhuis, Arjen P. de Vries, Maarten de Rijke, and Faegheh Hasibi. 2025. Adaptive Orchestration of Modular Generative Information Access Systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3899–3910. https://doi.org/10.1145/3726302.3730351

[37] Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. RouterBench: A Benchmark for Multi-LLM Routing System. In *Agentic Markets Workshop at ICML 2024*. https://openreview.net/forum?id=IVXmV8Uxwh

[38] Canbin Huang, Tianyuan Shi, Yuhua Zhu, Ruijun Chen, and Xiaojun Quan. 2025. Lookahead Routing for Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=DRIRD9ELMb

[39] Kung-Hsiang Huang, Akshara Prabhakar, Onkar Thorat, Divyansh Agarwal, Prafulla Kumar Choubey, Yixin Mao, Silvio Savarese, Caiming Xiong, and Chien-Sheng Wu. 2026. CRMArena-Pro: Holistic Assessment of LLM Agents Across Diverse Business Scenarios and Interactions. *Transactions on Machine Learning Research* (2026). https://openreview.net/forum?id=EPlpe3Fx1x

[40] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems* 31 (2018).

[41] Dietmar Jannach and Michael Jugovac. 2019. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)* 10, 4 (2019), 1–23.

[42] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 7036–7050. https://doi.org/10.18653/v1/2024.naacl-long.389

[43] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 383–390. https://doi.org/10.1145/3306618.3314288

[44] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=Rwhi91ideu

[45] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[46] Hideaki Joko, Shakiba Amirshahi, Charles L. A. Clarke, and Faegheh Hasibi. 2026. WildClaims: Conversational Information Access in the Wild(Chat). In *Advances in Information Retrieval*. Springer Nature Switzerland, Cham, 367–382.

[47] Jushaan Singh Kalra, Xinran Zhao, To Eun Kim, Fengyu Cai, Fernando Diaz, and Tongshuang Wu. 2025. MoR: Better Handling Diverse Queries with a Mixture of Sparse, Dense, and Human Retrievers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 11971–11990. https://doi.org/10.18653/v1/2025.emnlp-main.601

[48] Sora Kang, Andreea-Elena Potinteu, and Nadia Said. 2025. ExplainitAI: When do we trust artificial intelligence? The influence of content and explainability in a cross-cultural comparison. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3706599.3720222

[49] Evangelos Kanoulas, Panagiotis Eustratiadis, Yongkang Li, Yougang Lyu, Vaishali Pal, Gabrielle Poerwawinata, Jingfen Qiao, and Zihan Wang. 2025. Agent-centric information access. *arXiv preprint arXiv:2502.19298* (2025).

[50] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[51] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. 2025. AI Agents That Matter. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=Zy4uFzMviZ

[52] Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. 2025. LLM Economist: Large Population Models and Mechanism Design in Multi-Agent Generative Simulacra. https://doi.org/10.48550/arXiv.2507.15815 arXiv:2507.15815 [cs].

[53] Timothy L Keiningham, Bruce Cooil, Lerzan Aksoy, Tor W Andreassen, and Jay Weiner. 2007. The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet. *Managing service quality: An international Journal* 17, 4 (2007), 361–384.

[54] To Eun Kim and Fernando Diaz. 2025. Towards Fair RAG: On the Impact of Fair Ranking in Retrieval-Augmented Generation. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy) *(ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 33–43. https://doi.org/10.1145/3731120.3744599

[55] To Eun Kim and Fernando Diaz. 2026. LTRR: Learning To Rank Retrievers for LLMs. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*. Melbourne, VIC, Australia. July 20–24, 2026.

[56] Anton Korinek and Jai Vipra. 2025. Concentrating intelligence: scaling and market structure in artificial intelligence. *Economic Policy* 40, 121 (2025), 225–256.

[57] Carlos Lassance and Stéphane Clinchant. 2022. An Efficiency Study for SPLADE Models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2220–2226. https://doi.org/10.1145/3477495.3531833

[58] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474.

[59] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[60] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556* (2025).

[61] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. 2025. Robust Information Retrieval. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining* (Hannover, Germany) *(WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 1008–1011. https://doi.org/10.1145/3701551.3703476

[62] Shayne Longpre, Christopher Akiki, Campbell Lund, Atharva Kulkarni, Emily Chen, Irene Solaiman, Avijit Ghosh, Yacine Jernite, and Lucie-Aimée Kaffee. 2025. Economies of Open Intelligence: Tracing Power & Participation in the Model Ecosystem. *arXiv preprint arXiv:2512.03073* (2025).

[63] Eli Lucherini, Matthew Sun, Amy A. Winecoff, and Arvind Narayanan. 2021. T-RECS: A Simulation Tool to Study the Societal Impact of Recommender Systems. *CoRR* abs/2107.08959 (2021).

[64] Harry Markowitz. 1952. Portfolio Selection. *The Journal of Finance* 7, 1 (1952), 77–91. http://www.jstor.org/stable/2975974

[65] Sendhil Mullainathan and Richard H Thaler. 2000. *Behavioral Economics*. Working Paper 7948. National Bureau of Economic Research. https://doi.org/10.3386/w7948

[66] Thomas E Nelson, Zoe M Oxley, and Rosalee A Clawson. 1997. Toward a psychology of framing effects. *Political behavior* 19, 3 (1997), 221–246.

[67] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. 2021. FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, Vol. 53. ACM New York, NY, USA, 20–43.

[68] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2025. RouteLLM: Learning to Route LLMs from Preference Data. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=8sSqNntaMr

[69] OpenRouter, Inc. 2026. *OpenRouter*. https://openrouter.ai/ Accessed: 2026-02-04.

[70] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[71] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep Dyna-Q: Integrating Planning for Task-Completion Dialogue Policy Learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2182–2192. https://doi.org/10.18653/v1/P18-1203

[72] Michele Polo. 2018. *Entry games and free entry equilibria*. Edward Elgar Publishing.

[73] Michael E Porter. 1997. Competitive strategy. *Measuring business excellence* 1, 2 (1997), 12–17.

[74] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 132–148. https://doi.org/10.1007/978-3-031-88708-6_9

[75] Hossein A Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2025. Syndl: A large-scale synthetic test collection for passage retrieval. In *Companion Proceedings of the ACM on Web Conference 2025*. 781–784.

[76] William Rand and Roland T. Rust. 2011. Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing* 28, 3 (Sept. 2011), 181–193. https://doi.org/10.1016/j.ijresmar.2011.04.002

[77] Stephen A Rhoades. 1993. The herfindahl-hirschman index. *Fed. Res. Bull.* 79 (1993), 188.

[78] Kerry Rodden, Hilary Hutchinson, and Xin Fu. 2010. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2395–2398.

[79] Sherwin Rosen. 1981. The economics of superstars. *The American economic review* 71, 5 (1981), 845–858.

[80] Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 525–532.

[81] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2395–2400. https://doi.org/10.1145/3626772.3657957

[82] Alireza Salemi and Hamed Zamani. 2024. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 741–751.

[83] Renee Shelby, Fernando Diaz, and Vinodkumar Prabhakaran. 2025. Taxonomy of User Needs and Actions. https://doi.org/10.48550/arXiv.2510.06124 arXiv:2510.06124 [cs].

[84] Ian Soboroff. 2023. Overview of TREC 2023. In *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14–17, 2023 (NIST Special Publication, Vol. 1328)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec32/papers/overview_32.pdf

[85] Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. 2025. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12658–12666.

[86] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026. Kimi K2. 5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276* (2026).

[87] Brian Uzzi. 1996. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review* (1996), 674–698.

[88] Hal R. Varian. 2010. *Intermediate microeconomics: a modern approach* (8. ed ed.). Norton, New York, NY.

[89] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory*. PMLR, 25–54.

[90] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. https://doi.org/10.48550/arXiv.2411.04368 arXiv:2411.04368 [cs].

[91] xAI. 2025. *Grok 4.1 Model Card*. Model Card. xAI. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf Accessed 2026.

[92] Chen Xu, Clara Rus, Yuanna Liu, Marleen de Jonge, Jun Xu, and Maarten de Rijke. 2025. Fairness in information retrieval from an economic perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 4126–4129.

[93] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[94] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. 2025. {τ}-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=roNSXZpUDN

[95] Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. 2026. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[96] Robert B Zajonc. 2001. Mere exposure: A gateway to the subliminal. *Current directions in psychological science* 10, 6 (2001), 224–228.

[97] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[98] Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2687–2692.

[99] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating Classroom Education with LLM-Empowered Agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 10364–10379. https://doi.org/10.18653/v1/2025.naacl-long.520

[100] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2023. WildChat: 1M ChatGPT Interaction Logs in the Wild. https://openreview.net/forum?id=Bl8u7ZRlbM

[101] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. GPTSwarm: Language Agents as Optimizable Graphs. https://openreview.net/forum?id=uTC9AFXIhg

[102] Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. 2017. Online learning to rank in stochastic click models. In *International conference on machine learning*. PMLR, 4199–4208.