Report on The Search Futures Workshop at ECIR 2026

Leif Azzopardi
Microsoft/University of Strathclyde
Scotland
leifos@acm.org

Charles L. A. Clarke
University of Waterloo
Canada
claclark@gmail.com

Claudia Hauff
Spotify
The Netherlands
claudia@hauff.edu

Yubin Kim
Vody
USA
yubin@vody.com

Adam Roegiest
Zuva
Canada
adam@roegiest.com

Johanne R. Trippas
RMIT University
Australia
j.trippas@rmit.edu.au

Zhaochun Ren
Leiden University
The Netherlands
z.ren@liacs.leidenuniv.nl

Saber Zerhoudi
University of Passau
Germany
szerhoudi@acm.org

Qingyao Ai, Shakiba Amirshahi, Marcel Gohsen, Jaap Kamps, Jussi Karlgren, Yiqun Liu, Shuo Miao, Behnaz Nojavanasghari, Jingfen Qiao, Naren Ramakrishnan, Eunice Son, Yiteng Tu, Suzan Verberne, Yumeng Wang, Chen Xu, Raquib Bin Yousuf, Jujia Zhao*

Abstract

The Third Search Futures Workshop [Azzopardi et al., 2026], in conjunction with the *Forty-eight European Conference on Information Retrieval (ECIR)* 2026, looked into the future of search to ask questions such as:

- *How can we navigate data privacy in large language model (LLM)-based information retrieval (IR)?*
- *How can we implement agentic IR for proactive knowledge synthesis?*
- *How do we ensure trustworthy information access beyond citations in the age of language models?*
- *How does deep search transition from matching to reasoning?*
- *What is meant by information semantics, knowledge representation, and natural language in a world of LLM-powered search?*
- *What are serendipity engines, and how do they explore proactive web search via LLM agents, retrieval augmented generation (RAG), and simulated user feedback?*

*Affiliation not shown for all authors due to space limitations (see Appendix A for details).

The third edition of the workshop opened with ten lightning talks from a diverse group of speakers. Rather than traditional paper presentations, these short talks offered concise overviews of emerging ideas and critical insights, enabling a rapid exchange across various topics. The format was designed to spark discussion and expose participants to a broad spectrum of future-facing research directions in a compact timeframe. This report, co-authored by the workshop organizers, presenters, and participants, summarizes the talks and key discussions. Our aim is to share these insights with the broader IR community and help seed further dialogue around the themes raised.

Date: 2 April 2026.

Website: https://searchfutures.github.io/.

1 Introduction

The *Thirds Search Futures Workshop*,[1] held in conjunction with ECIR 2026,[2] provided a platform to explore and debate the future of search. To support collaborative discussion, presentation slides from both invited talks and breakout sessions were openly shared, enabling participants and the wider community to engage and contribute comments via an interactive forum [3].

The Third Search Futures Workshop continued the conversation from previous editions [Azzopardi et al., 2024a,b; Clarke et al., 2025a,b], bringing together researchers, practitioners, and designers to critically examine the evolving landscape of search in the age of generative AI. With growing momentum in LLMs and emerging applications that challenge traditional paradigms, the workshop aimed to refine and expand our understanding of what search could — and should — become.

This series was initially inspired by discussions at ACM SIGIR 2023, where the generative AI revolution prompted a central question: "Is information retrieval still relevant?" That question drove the first workshop, which created space to reflect on the potential futures of search, considering both the strengths and threats posed by these technologies and their implications for end-users, system designers, researchers, and society.

The *Second Search Futures Workshop* was shaped by pressing questions about the future of search in the age of generative technologies. As in the first edition, participants engaged with concerns such as: *How can we trust Generative IR? What is the role of search when content can be generated on demand? How do we distinguish fact from fiction? Could these tools steer us toward the dystopias imagined in science fiction?*

However, despite these concerns, the workshop maintained a tone of thoughtful optimism. Across lightning talks and breakout sessions, participants proposed new applications, methodological innovations, and design principles to reshape IR constructively. The discussions also revisited deeper questions about the foundations of the field itself: *What does IR stand for? What values and principles should guide us going forward? What do we do with all the uncertainty in the world and information?* [Rieger et al., 2026] In addition, we discussed the major themes discussed in the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025) [Trippas and

[1] https://searchfutures.github.io/

[2] https://ecir2026.eu/

[3] https://docs.google.com/presentation/d/1inTbDUPsngKZ86kxPzVRtP-gEEOLsVUEDtfIp31vEwc/edit?usp=sharing

Culpepper, 2025] for cross-pollination between current topics. The participants at the ECIR workshop contributed to lively exchanges, and the second workshop emphasized the challenges ahead and the many emerging opportunities and open research questions that will define the next era of IR.

2 Vision Statements

During the workshop, speakers shared their perspectives on the future of search. The following statements provide a summary of their viewpoints in their own words. For presentation here, the statements are listed alphabetically by first author. During the workshop, talks were grouped around applications, theoretical perspectives, and methodological innovations, each engaging with the opportunities, challenges, and implications for users, society, and IR.

Feed Your Need by Information Farming

Leif Azzopardi and Adam Roegiest

IR, as a field, was built on the assumption that information exists "out there" and that the user's task is to find it. From ranked retrieval to Berry Picking and Information Foraging Theory, decades of research have refined how people navigate information landscapes, evaluate patches, and accumulate relevant pieces under conditions of scarcity, cost, and uncertainty. But this assumption is rapidly collapsing. With generative AI, users are no longer primarily finding information. They are growing it!

Information Farming names this shift [Azzopardi and Roegiest, 2026]. Instead of issuing queries and traversing sources, users plant prompts, cultivate outputs through iteration, and harvest structured information within generative systems. Search effort moves from navigation to cultivation; uncertainty moves from the environment into the model; and information needs are fed through production rather than discovery. If this trajectory continues, then much of what we currently call "search" risks becoming a legacy interaction pattern that is no longer the dominant modality. In this talk, we argue that Information Farming is evolution of Information Foraging. When information is grown rather than found, the core problems that IR research has historically addressed (i.e., ranking, coverage, query formulation, patch selection), are no longer sufficient. Feeding information needs through farming, however, introduces new challenges that demand a reorientation of the field.

The first challenge is the disappearance of epistemic friction. Traditional search externalizes doubt: conflicting sources, incomplete coverage, and visible provenance force users to compare and evaluate. Farming collapses these frictions into a single, fluent output. Errors, hallucinations, and bias are no longer encountered as "bad sources" but as properties of the crop itself. The danger is not that systems are sometimes wrong, but that they are convincingly wrong. Future IR systems must therefore confront how trust, uncertainty, and accountability are surfaced when there is no obvious "source" to inspect.

The second challenge is who gets to eat well. Information Farming reallocates effort away from exploration and toward prompt design, iterative refinement, and workflow construction. These skills are unevenly distributed and poorly supported by current systems. As with early agriculture,

farming creates surplus, but also inequality. Those who know how to cultivate effectively can feed their information needs cheaply and reliably; others remain dependent on brittle or low-quality crops. Search futures must grapple with whether generative systems democratize access to knowledge or entrench new forms of informational privilege.

The third challenge is ecological collapse at scale. Farming encourages reuse: prompts are replanted, outputs recycled, and generated content propagated into downstream systems. Without intentional diversification, this creates mono-cultures, homogeneous knowledge, amplified biases, and self-reinforcing feedback loops. Unlike foraging, where diversity emerges from environmental heterogeneity, farming demands active stewardship. The future of IR may depend less on better ranking and more on maintaining healthy information ecosystems. If search is no longer about finding documents, but about feeding information needs over time, then the field must confront uncomfortable questions:

- What is "relevance" when information is generated rather than retrieved?
- How should IR systems expose uncertainty, provenance, and disagreement in farmed outputs?
- What literacies do users need to become competent information farmers, and how should systems support them?
- When does reuse become contamination, and how do we prevent informational monocultures?
- If farming becomes the dominant mode, what remains of search—and what should replace it?

We do not argue that foraging disappears. Rather, we argue that search futures must be designed for a world where foraging is no longer the default. If we continue to optimize search systems for users who roam information landscapes, while users increasingly stay put and grow what they need, we risk perfecting tools for a problem that is already fading. The future of search may not lie in helping users look harder—but in helping them grow better.

Evaluation Paradigm of the Future? A Case for Valid, Fair and Reproducible Evaluation with User Simulation

Marcel Gohsen

The popularity of the conversation-based interaction paradigm for accessing information is steadily growing, so that it should only be a matter of time before conversational search becomes the dominant mode to search for information. Yet, the question of how the effectiveness of conversational search systems can be evaluated remains open. User simulation emerged as a promising paradigm for the evaluation of conversational search systems. However, a central question for user simulation is still unanswered: what characteristics must such a simulator have in order to produce "valid" assessments? Currently, the field of user simulation lacks standards to validate the simulators themselves and thus to guarantee that system evaluation results accurately reflect the system performance.

Researchers in the user simulation community are often concerned with how "realistic" or how "authentic" the produced utterances of their simulators are in order to validate the produced simulations. However, for the practical application of user simulation for system evaluation, the

necessary extent to which utterances have to be realistic is not yet clear. We propose that three essential characteristics must hold for a useful user simulation. A user simulation should deliver the same system performance rankings as if the evaluation had been performed with real users (validity). Furthermore, if the simulation-based evaluation is repeated, the same system performance rankings should be obtained with certain confidence (reproducibility). Finally, simulated users should ask queries of similar difficulty to all systems (fairness).

In an effort to develop user simulation approaches and standard methodology to assess the simulators, the TREC 2026 User Simulation[4] track was initiated. This track represents a long-term effort to establish user simulation standards and to evolve this evaluation paradigm into a standard tool for the development of conversational search systems through collaborative efforts of the IR community. However, it will take time until the shared task will reach that point. In the meantime, we derived immediate best-practices from the practical deployment of user simulation in three shared tasks (Touché RAD'25 [Kiesel et al., 2025], TREC iKAT'25 [Aliannejadi et al., 2025], and the micro-shared task at Sim4IA'25 [Schaer et al., 2025]) in order to facilitate the development of user simulators as reusable resources [Gohsen et al., 2025]. Collaboratively exploring user simulation in shared tasks as a unified community will lead to a future in which user simulation replaces the Cranfield-paradigm for more meaningful evaluations of conversational search systems.

Trustworthiness as a feature to address in system design and as an educational challenge for media literacy

Jussi Karlgren

Information access and search, for the most general use cases, is moving to conversational question answering. This implies a new set of evaluation challenges, where the qualities of the interaction and the conversation itself risk obscuring the quality of the underlying search engine [Bauer et al., 2025].

One of the most crucial quality criteria of an information service is that of *trust*, and "trustworthiness" is repeatedly brought up by those who wish to regulate the offerings of information services based on artificial intelligence.[5]

Trustworthiness is not a technical quality of the model or of the system. Trust cannot be built into a system a priori or regulated after the system is put into service. Trust is a social feature, and it is earned and lost through track record and reputation. This does not mean that the design of the service is irrelevant to trust. A snazzy look will give a good impression and potentially inspire trust; a well-designed interaction will enable the system to project trustworthiness better than a shoddily assembled one would. For an interactive conversational system, a defining design feature is that of conversational linguistic style, and indeed much of the success of some established web search engines were based on interaction design, moving from cluttered portals to dedicated search pages with a promise of independence from outside influence.

[4] https://usersim.ai/trec/

[5] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government
https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2022/0901/Pages/TrustworthyAI.aspx

Conversational and behavioural principles are global on one level: general principles of politeness and mindfulness, e.g., are accepted by most everyone; their implementation varies in interesting and entertaining ways across cultures and linguistic areas [Karlgren and Sahlgren, 2025]. **This calls for information system designers to put serious effort into aligning conversational behaviour to fit cultural expectations and what the goals for information provision are.** This is not routinely done today for conversational interfaces built on generative artificial intelligence: instruction training data sets are automatically translated from North American West Coast English into target languages, without cultural localisation.

Behavioural and conversational cues help us make sense of a counterpart's communicative goals and help us figure out if someone is trustworthy or not. Such cues can be exploited for nefarious purposes [Mitnick and Simon, 2003]. If an interactive conversational information system projects trustworthiness through its design it can be tacked on to information services with deceptive goals. **This calls for requiring instruction training of conversational systems to be done transparently as a matter of course, in order to allow for an open audit of system qualities.** Such requirements cannot easily be regulated into place, but by trademarking conversational functionality the profession will be able to establish appropriate codes of practice.

Our educational systems prepare coming generations to evaluate the quality of static written material: practising reviewing, summarising, referencing textual material takes up a considerable portion of the school curriculum. For live conversations, in parallel, conventional wisdom gives us advice on how to handle sketchy counterparts: parents and other more wary peers help us not be entrapped by con artists. Still, occasionally to all of us end up fooled. Cloaking an unreliable knowledge source in a well designed conversational format risks fooling some users some of the time. For non-textual and dynamically generated material, conventions on how to decode persuasion are not as well-established and schools have no readiness to address trustworthiness in information systems as systematically as they discuss written propaganda or news. **This calls for us as information system designers to provide the world in general, and educational systems specifically, with conventions or pointers on how to understand, assess, and react to persuasive generative artificial intelligence output.** This will be a valuable addition to the current field of *Media literacy*.

From Search to Synthesis: Retrieval Beyond Similarity

Jingfen Qiao

LLMs have shifted knowledge acquisition from search to synthesis, yet RAG systems remain essential for grounding facts and mitigating hallucinations. Traditional RAG architectures typically isolate Planner, Executor, and Critic modules [Yao et al., 2023; Asai et al., 2024]. This separation creates a capability asymmetry. A strong Planner such as GPT-4 reasons at a high level of abstraction, while a weak Retriever based on embedding similarity cannot interpret such instructions. The Retriever returns noisy outputs that the Planner cannot effectively filter. This decoupled design leaves the asymmetry unresolved, preventing system-level coordination.

We propose reframing retrieval as dynamic exploration. Complex queries require iterative refinement based on intermediate results. This process suits sequential decision-making. We model retrieval as a Markov Decision Process (MDP) [Jin et al., 2025; Chen et al., 2026; Zhang et al., 2026]. States encode accumulated evidence. Actions include exploring new sources, backtracking

from unproductive paths, or exploiting current evidence to formulate an answer. This formulation shifts the objective from maximizing recall to optimizing the reasoning trajectory. Integration with Monte Carlo Tree Search (MCTS) [Hao et al., 2023] enables systematic exploration of the retrieval space.

Within this MDP framework, a Bridge Agent coordinates between the Planner's high-level policy and the tool ecosystem. The Planner decides when to explore, backtrack, or exploit. The Bridge Agent executes these decisions through three functions. First, Query Translation converts strategic instructions into DSL or API calls. Second, Dynamic Routing selects appropriate tools such as vector databases, real-time APIs, or relational databases [Schick et al., 2023]. Third, Information Filtering processes retrieved results to remove noise before updating the MDP state.

Next-generation retrievers should exhibit three characteristics. First, sensitivity to planning intent allows them to interpret strategic goals as executable queries rather than keyword lookups. Second, robustness in handling history enables them to leverage past trajectories when making decisions rather than treating each retrieval independently. Third, capacity to actively explore empowers them to backtrack and self-correct through learned policies rather than fixed heuristics.

What Should Search Retrieve Now?

Eunice Son, Raquib Bin Yousuf, and Naren Ramakrishnan

New integrations of LLMs into search systems are constantly being proposed and this has caused soul-searching in IR circles. If content can be generated, paraphrased, or recombined at scale, what exactly should retrieval systems retrieve, and on which signals should they rely? In environments where textual content is noisy, weakly aligned with user queries, or even partially machine-generated, reliance on document text alone may no longer be sufficient.

We argue that, in today's generative era, search systems will need to retrieve more than content. They will need to treat schema-aware signals as first-class retrieval evidence. By schema-aware signals we mean, for instance, metadata attributes, field semantics (units, time periods, geography, cohorts), relational structure, and authorship information. These signals can act as semantic and provenance anchors: they help the system stay grounded in who/what the information is about, when it applies, and where it came from, even when surrounding language is rewritten or noisy.

We further motivate this shift by observing two increasingly common failure modes of text-centric retrieval. First, generative content floods make it difficult to distinguish authoritative from synthetic content using embeddings alone. Second, in structured corpora, language is reused across entities and time periods, making chunk-level similarity insufficient for disambiguation. Two examples of such corpora are, U.S. Securities and Exchange Commission (SEC) filings and the U.S. Integrated Postsecondary Education Data System records (IPEDS).

In both of the aforementioned cases, the relevant signal resides not in the prose, but in the schema-level organization of the data. For instance, consider a user exploring IPEDS, where the user asks qualitative questions pertaining to higher education enrollment trends. Successful query answering requires focusing on specific institutions, time periods, and attributes. A second example can be the queries into the SEC filings. In that setting, identical or near-identical language appears across documents, and without metadata such as company, year, or section, retrieval systems can return plausible but incorrect evidence.

This challenge becomes more pronounced in emerging recursive and agentic retrieval settings. In architectures such as Recursive Language Models (RLMs) [Zhang et al., 2025a], a root model decomposes a query into sub-tasks and delegates them to sub-models. However, schema information does not automatically propagate across these recursive calls. As a result, even when schema improves reasoning at the root level, its benefits are lost unless explicitly transmitted. This suggests that retrieval must support not only what is retrieved, but also what flows across reasoning steps, including schema, rules, and task context.

One core idea of improvement is to "verbalize" records using templates that explicitly name the fields, turning a record into a canonical textual form Dinh et al. [2022]. In a retrieval setting, the same approach can make schema signals retrievable by design: instead of hoping an embedding model infers entity/time/attribute from a noisy representation, we can index sentence forms that explicitly contain those anchors. Another axis is how much information is necessary to surface and disambiguate correct evidence [Chen et al., 2025b; Anthropic, 2026], which raises questions about minimal yet sufficient schema representations. Early observations suggest that simple signals such as entity and time provide strong disambiguation, while additional schema may yield diminishing returns.

Such schema-aware retrieval naturally connects to trust and accountability in generative search. As generated content proliferates, schema-aware signals can support provenance tracking, entity grounding, and temporal alignment. This reframes retrieval not just as matching content, but as recovering verifiable evidence tied to the correct entity, time, and source.

We outline several open questions. How should structure be represented for neural retrieval? How should schema signals be indexed efficiently without frequent re-indexing? What should be propagated across recursive reasoning pipelines? And how should we evaluate success when correctness depends on provenance rather than textual overlap? Existing benchmarks measure content relevance, but schema-aware retrieval requires evaluation along dimensions such as entity correctness, temporal alignment, and traceability.

This vision also connects to parallel work in the space of data exploration systems, where users interact with large structured datasets through schema-aware interfaces and agentic assistance [Chen et al., 2025a; Sahu et al., 2024]. These settings highlight that many real-world search tasks involve navigating structured information spaces rather than locating a single document. This suggests that future retrieval models may need to reason jointly over text and structure, moving from document retrieval toward evidence retrieval grounded in schema.

Analytical Search

Yiteng Tu, Shuo Miao, Yiqun Liu, and Qingyao Ai

Over the past decades, IR has been highly successful in helping users address information finding problems: given a query, retrieve topically relevant documents to satisfy their needs [Ai et al., 2023; Metzler et al., 2021; Shah and White, 2025]. However, as IR systems are increasingly applied in domains such as finance, law, scientific research, and policy analysis, a growing class of user needs beyond locating relevant documents or answering isolated factual questions has become evident [Luo et al., 2025; Hu et al., 2025; Sun et al., 2025a; Singer et al., 2013]. With such **analytical information needs**, users are not limited to asking "*What is X?*" but rather "*How many cases meet the requirements?*" or "*What statistical patterns exist?*". Answering such

questions requires aggregating information across multiple sources, aligning evidence along varying dimensions, and synthesizing conclusions through multi-step reasoning.

Classical IR systems [Robertson et al., 2009; Karpukhin et al., 2020] primarily leave such analytical burden to users. They handle it through labor-intensive workflows involving repeated searching, manual filtering, validation, and reasoning. Recent advances in LLMs [Hurst et al., 2024; Guo et al., 2025; Yang et al., 2025; Touvron et al., 2023] have given rise to the impression that these challenges may be solved by simply coupling retrieval with generation. While LLMs excel at fluent language generation and can perform impressive reasoning in constrained settings, existing "IR+LLM" systems remain fundamentally limited for analytical tasks [Luo et al., 2025; Hu et al., 2025]. They typically treat retrieval as a supporting mechanism for answer generation, rely on a small set of top-ranked documents optimized for topical relevance, and provide little control over evidence selection, reasoning structure, or verifiability. As a result, they struggle with analytical queries that demand high recall, structured reasoning, and accountability.

To this end, we propose **analytical search** as a distinct search paradigm. Fundamentally, it reframes search not as document ranking or answer generation, but as an evidence-governed analytical process. Its objective is not to return documents or produce fluent answers, but to generate justified conclusions grounded in explicitly retrieved and verifiable evidence, which leads to several defining characteristics. First, analytical search is conclusion-oriented rather than answer-centric. Second, it adopts a notion of complex relevance based on the utility in reasoning rather than topical similarity. Finally, analytical search is inherently evidence-governed: every claim or inference should be traceable to explicit sources, enabling inspection, validation, and auditing.

From a system perspective, solving analytical queries can be viewed as an end-to-end analytical workflow, involving query understanding and decomposition, recall-oriented evidence retrieval, reasoning-intensive fusion, and adaptive verification. However, the significance of analytical search lies less in any specific architecture and more in the conceptual redefinition of what search optimizes: from relevance to reasoning, from answers to conclusions, and from information access to analytical problem-solving.

It is important to distinguish the analytical search system from several related paradigms. Compared to retrieval-augmented generation (RAG) [Borgeaud et al., 2022; Su et al., 2024; Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021], analytical search does not treat retrieval as context for generation, but as the construction of an evidence base for reasoning. Compared to deep research [Huang et al., 2025; Zhang et al., 2025b; Li et al., 2025], which emphasizes exploration and narrative synthesis, analytical search focuses on well-defined analytical questions with clear termination criteria: when sufficient evidence has been gathered, and justified conclusions can be reached. Finally, unlike agentic databases [Hu et al., 2025; Sun et al., 2025a; Tang et al., 2025], which operate primarily over structured data and deterministic query execution, analytical search must integrate heterogeneous sources retrieval and reasoning in a dynamic open-world setting.

In summary, analytical search represents a promising and necessary direction for the future of IR. As an increasing number of users turn to IR systems for complex analytical problems, the IR community should move beyond relevance ranking and answer generation, embracing analytical search as a structured, evidence-based reasoning process.

Rethinking Ranking in the Latent Space: Activation-Level Control for the Future of Search

Yumeng Wang and Suzan Verberne

Recent studies have shown that LLMs can function as effective zero-shot rankers [Liang et al., 2022; Qin et al., 2023; Zhuang et al., 2024; Sun et al., 2023], yet their ranking performance is highly sensitive to prompt formulation [Sun et al., 2025b]. In particular, role-play prompting [Shanahan et al., 2023] – assigning the model a functional role such as a "careful" or "careless" search assistant – can lead to large and sometimes unexpected shifts in retrieval quality [Wang et al., 2025]. Mechanistic analyses using causal intervention techniques, such as activation patching [Meng et al., 2022; Vig et al., 2020], provide insight into this sensitivity. These studies suggest that role-related signals are encoded early in the network and propagate through activation pathways that interact only weakly with query–document representations.

Building on this perspective, we recently explored whether ranking behavior can be directly controlled at the activation level [Wang et al., 2026] with activation steering methods [Zou et al., 2023; Bartoszcze et al., 2025]. Rather than modifying prompts or finetuning models, this line of research identifies latent directions corresponding to different ranking factors and intervenes on hidden representations at inference time. Empirical results show that such activation steering, can substantially improve pointwise ranking using only a small number of anchor queries, without changing model weights or introducing explicit cross-document comparisons. These improvements suggest that a significant amount of ranking capacity in LLMs is present but under-utilized, and can be recovered through post-hoc calibration of internal representations.

More importantly, this emerging view reframes ranking as a latent control problem. Geometric analyses [Wang et al., 2026] reveal that, for a given query, document representations often form simple, low-dimensional structures in latent space. Activation steering does not invent new ranking behaviors, but reshapes how existing representations are expressed along these structures, including reducing dispersion, stabilizing ordering, and improving consistency. From this perspective, we see that ranking errors can arise both from missing relevance knowledge and from how internal signals are scaled, combined, and expressed during inference.

This latent-space view of ranking opens up new possibilities for the future of search and recommendation systems. Because activation-level interventions are lightweight, reversible, and do not require retraining/finetuning, they enable rapid adaptation of ranking behavior to changing objectives. In particular, personalization can be naturally formulated as steering: user preferences, risk tolerance, or domain-specific priorities can be encoded as latent directions and applied on a per-user or per-session basis. Unlike traditional fine-tuning or reinforcement learning approaches, such adaptation can be performed dynamically, with minimal data and without maintaining separate models.

More broadly, this paradigm suggests a shift away from monolithic ranking models toward configurable rankers, where a shared pretrained backbone supports multiple ranking behaviors through latent control. Such systems could support personalization, domain adaptation, and policy constraints in a unified framework, offering a promising direction for the next generation of search and retrieval systems.

Toward a New Economic Ecosystem in Generative Information Retrieval

Chen Xu

This dissertation studies fairness-aware re-ranking through the lens of attention economics. In this view, ranking systems form an attention market: users provide limited attention, platforms allocate exposure through ranking positions, and suppliers compete for visibility. Prior chapters show that this market can be regulated through economic mechanisms [Xu et al., 2023, 2024, 2025b,a]. These works suggest a common principle: fairness in ranking can be understood as the redistribution of scarce attention through controllable market instruments.

Generative IR changes the underlying economy. In RAG and generative search, users no longer consume a ranked list directly; instead, they consume a synthesized answer produced from retrieved evidence [Lewis et al., 2020; Liu et al., 2023]. Therefore, the key scarce resource is no longer only ranking exposure. It also includes input-side context tokens, which determine which documents or passages enter the model, and output-side attribution tokens, which determine which sources are cited, quoted, or credited in the final answer. This shift turns the traditional attention market into a token-mediated market.

This new market also changes the role of suppliers and users. Suppliers are no longer merely competing for ranked positions; they compete to be selected into the context window, preserved during generation, and credited through citations. Users are no longer browsing and comparing many ranked results; they often rely on the generated answer and its citations as a compressed interface to knowledge. Generative engine optimization (GEO) is a concrete example of this transition [Aggarwal et al., 2024]: content providers strategically rewrite pages to increase their visibility in generative engines, turning citation probability into a new economic objective. In this sense, GEO is not simply a new version of SEO, but a supplier-side optimization strategy in a token-based allocation system.

This perspective also opens a broader research agenda that connects token economics with game theory and multi-agent systems. Generative IR is not a static allocation problem, but a repeated strategic interaction among suppliers, platforms, and users. Suppliers may adapt their content to increase context inclusion and citation probability; platforms may adjust retrieval, attribution, and verification mechanisms; users may shift trust and behavior according to generated answers.

This naturally suggests game-theoretic formulations, such as Stackelberg games for platform–supplier interaction, evolutionary games for repeated content adaptation, and mechanism design for incentive-compatible attribution [Fudenberg and Tirole, 1991; Taylor and Jonker, 1978]. At the same time, recent LLM-based multi-agent systems provide a practical tool for simulating such ecosystems, where different agents can play the roles of creators, platforms, verifiers, advertisers, and users [Chen et al., 2024; Hao et al., 2026]. This makes it possible to study not only one-shot fairness in ranking or generation, but also long-term strategic effects such as citation competition, supplier adaptation, platform defense, and equilibrium welfare. In this sense, future fair generative IR systems should be viewed as economic institutions: they must allocate tokens and attribution efficiently, while also shaping incentives so that strategic suppliers are rewarded for producing verifiable, high-quality, and socially valuable information.

Toward Unified and Personalized Information Access with LLM-based Search and Recommendation

Jujia Zhao, Suzan Verberne, and Zhaochun Ren

Search and recommendation are two primary pathways through which users access information online [Zhang et al., 2024; Wu et al., 2024]. Search allows users to actively express immediate information needs through explicit queries, while recommendation systems infer long-term interests from historical interactions and proactively surface relevant items [Xie et al., 2024; Lin et al., 2026]. Although these two paradigms differ in interaction form, they are fundamentally connected: both aim to understand user intent and satisfy information needs [Yao et al., 2021]. This shared objective suggests that search and recommendation should not be treated as isolated systems, but as complementary observations of the same underlying user information need [Shi et al., 2024].

The emergence of LLMs provides a timely opportunity to revisit this separation [Lin et al., 2024]. LLMs offer strong semantic understanding, contextual reasoning, and generative capabilities, making them a natural foundation for modeling queries, user histories, item semantics, and behavioral signals within a unified framework [Zhao et al., 2026]. In principle, such a framework could enable search to become more personalized through long-term user preference modeling, while recommendation could become more responsive to users' immediate intents expressed through search behavior. This points toward a broader vision of unified and personalized information access, where users interact with a system that understands both what they explicitly ask for and what their historical behaviors imply.

However, realizing this vision requires more than simply applying existing LLM fine-tuning techniques to search and recommendation data. The two tasks emphasize different optimization signals. Search focuses on query-item relevance under a specific short-term intent, whereas recommendation relies more heavily on long-term user preferences, sequential patterns, and collaborative behavioral signals. When these objectives are optimized jointly in a shared LLM, their gradients may interfere with each other, leading to unstable training and negative transfer across tasks. At the same time, task-specific adaptation can disturb the pretrained semantic knowledge of the LLM, weakening its ability to interpret user intent, constraints, and contextual signals. These issues suggest that unified LLM-based search and recommendation is not merely a question of model capacity or parameter efficiency, but also a question of how to coordinate heterogeneous optimization signals while preserving the general knowledge that makes LLMs useful in the first place.

This position highlights two central challenges for LLM-based unified search and recommendation: cross-task optimization conflict and preservation of user intent understanding. A promising direction is to move beyond direct parameter sharing or naive parameter-efficient fine-tuning, and instead design optimization mechanisms that explicitly separate shared and task-specific signals, reduce destructive gradient interference, and constrain task adaptation to avoid unnecessary disruption of pretrained semantic knowledge. Such mechanisms would allow LLMs to absorb both search and recommendation signals without forcing them into a single undifferentiated update space.

From this perspective, unified search and recommendation should be viewed as a step toward the next generation of personalized information access. The goal is not only to improve ranking metrics on two related tasks, but to build systems that can jointly reason over explicit queries,

implicit preferences, historical behaviors, and item semantics. By identifying the optimization and knowledge-preservation challenges behind this goal, this work argues for a more principled approach to adapting LLMs for unified information access: one that treats search and recommendation as complementary user-intent signals and seeks to coordinate them within a stable, scalable, and deployment-friendly modeling framework.

3 Breakout Group Summary

In previous editions of the workshop [Azzopardi et al., 2024a; Clarke et al., 2025b], breakout discussions were convened around tables in the conference venue (and, on one occasion, on a bus to the airport). Small groups of attendees self-organised around a topic, talked for an hour or two, and a rapporteur produced a written summary after the fact. For ECIR'26, we experimented with a complementary format. Rather than convening physical breakout groups, we used the *Panel framework* of **OpenIIR**[6], an open platform for multi-agent IR simulation [Zerhoudi, 2026], to instantiate four simulated panels. Each panel was defined by a distinct theme and associated world model, and all four drew on a shared set of eleven LLM-driven persona agents. The four themes were drawn from a set of topics nominated by the workshop participants at the start of the breakout section, and grouped by the organisers into the four panels described below. The persona biographies were drafted on the basis of each participant's recent papers and talks, then reviewed and approved by the workshop participants before the panel commenced.

In practice, setting up a panel in OpenIIR involves three elements: the world model the panel will reason over, the personas that populate the panel, and the discussion structure. A *persona* is represented by a single biographical paragraph that describes the participant's name, affiliation, and the positions they have publicly taken in recent writing and talks. For the workshop, we rewrote each bio for each theme to capture where that participant was most likely to sit on the question the panel would discuss. Personas can be authored by hand or automatically generated from the world model via an extraction pass provided by the platform. Each persona is also assigned a token budget and a web search budget, which are debited as the panellist speaks and issues queries. Once a persona's token budget is exhausted, the moderator skips that persona on subsequent turns, which prevents any single panellist from monopolising the discussion.

For each panel, we built a separate *world model* using Gemini Deep Research. For each theme, we issued a deep-research query, then seeded the panel with the resulting synthesised report and all of its cited sources, including research papers as PDFs and web articles converted to Markdown. OpenIIR splits the world model into chunks and exposes it through a keyword-lookup tool that the panellists can call freely. If a panellist needs information outside the world model, they can issue a web search, but only after providing a written justification; the system rejects under-justified queries before any search budget is debited. World models are isolated to each panel, keeping each discussion grounded in the sources for its own theme.

The discussion is organised into named rounds, each defined by a name, a goal sentence, and a hard cap on the number of turns the round may take. Three of the four panels followed a familiar *opening*, *deliberation*, *wrap-up* structure: the opening elicits concrete positions, the deliberation works through a small set of pre-agreed sub-questions, and the wrap-up forces each panellist to

[6] https://openiir.com

commit to an open problem or a named research direction. Theme 3, the only generative session of the four, used a different structure appropriate to that task: a *pitch storm*, in which each panellist proposed several startup ideas; an elimination round, in which each idea was attacked by the panellist sitting to the speaker's left; and a final ranking round. We describe that format in detail in Section 3.3. Within any round, an LLM-driven moderator agent decides at each step who speaks next and what they should address. The same moderator can also advance the discussion to the next round or end the panel entirely once the round's goal has been met.

After every utterance, a separate *extractor* reads the new text, tags any new claims with their stance towards the topic (supporting, challenging, or neutral), and links those claims back to earlier ones as *supports*, *counters*, *refines*, or *questions*. For instance, an utterance asserting that "Q-RAG embeddings paper over a deeper planning problem" is recorded as a *neutral* claim that *counters* an earlier *supporting* claim that "Q-RAG is the right abstraction for long-context multi-step retrieval". The structure of claims and links that emerges is what we refer to as the *argument graph*; the chronological log of what was said in what order is the *transcript*. We used the argument graph as a sanity check that the synthesised report captured the major disagreements in the discussion, not just the points on which the panellists converged. Each simulation also exposes a separate persona ontology, an agent-roster view, a synthesised report, and a post-hoc interrogation surface on which one can either chat with a single panellist (whose context is restricted to their own utterances) or pose the same question to all panellists at once and compare their responses. Each panel ran to roughly thirty utterances spread across the three rounds and finished in about fifteen minutes.[7]

A practical consequence of this format, one that the in-person breakouts do not permit, is that every participant could be in every discussion at the same time. Rather than picking a single table to sit at, each participant was effectively simulated across all four panels, engaging with each theme through their own persona.

The persona biographies were drafted from detailed profiles of each participant produced by Gemini 3.1 Pro on the basis of their recent papers, talks, and other publicly available online sources. The eleven participants whose profiles seeded the panels were Shakiba Amirshahi (University of Waterloo), Charles Clarke (University of Waterloo), Marcel Gohsen (Bauhaus-Universität Weimar), Claudia Hauff (Spotify), Jaap Kamps (University of Amsterdam), Yubin Kim (Vody), Behnaz Nojavanasghari (Carnegie Mellon University), Jeremy Pickens (Elevate), Adam Roegiest (Zuva), Yiteng Tu (Tsinghua University), and Saber Zerhoudi (University of Passau). Each gave explicit consent for their persona to be used in both the simulations and this report.

3.1 Breakout 1: Analytical Search and Deep Research Agents

The agents simulating the eleven panellists were asked when a 20-turn analytical-search trajectory beats one well-designed retrieval, and how to evaluate the answer.[8] Each agent opened with a recent task pulled from the bio it was conditioned on. The shape of the resulting discussion is summarised in Figure 1.

[7]The panellists, the report synthesis pass, the moderator and the per-utterance claim extractor were powered by `gpt-5.4`.

[8]Full simulation, including transcript, argument graph, persona ontology, agent roster, generated report, and per-panellist interrogation surface: https://openiir.com/simulation/sim_0d9f423da0ac.

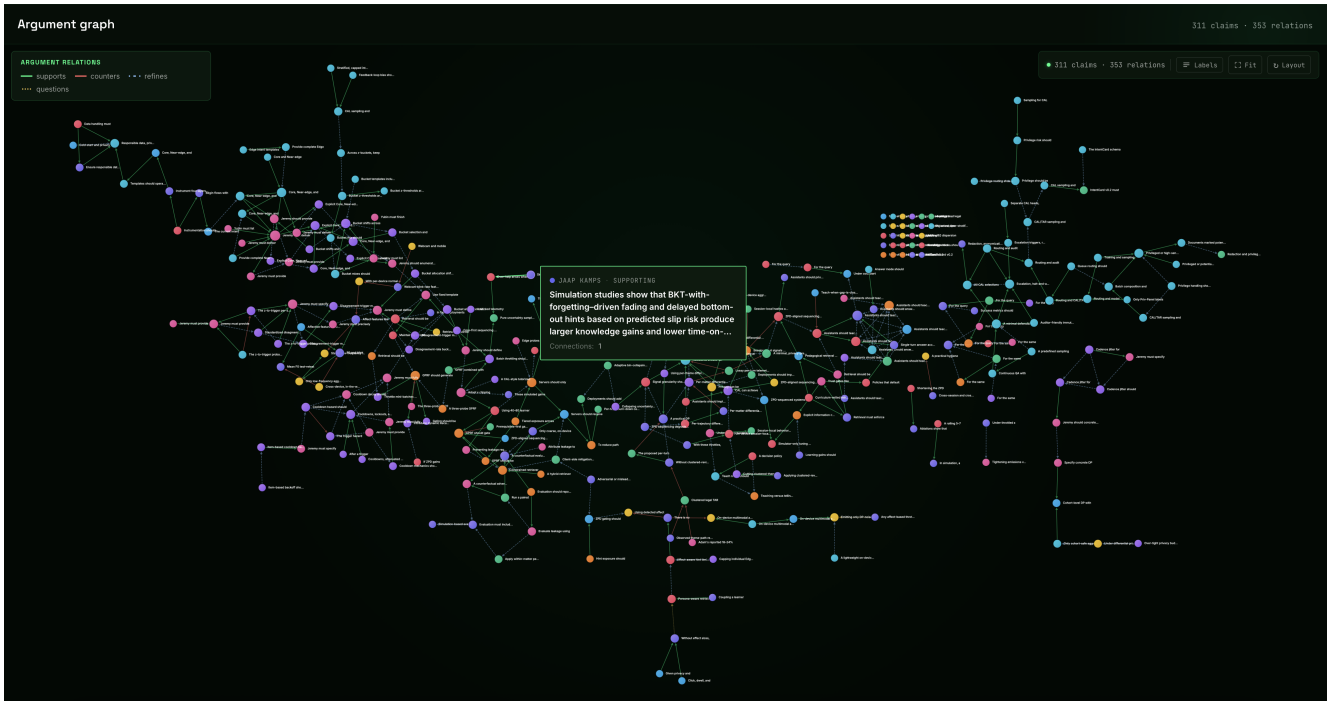

Figure 1. Argument graph of the Breakout 1 panel. Each node is a claim made by one of the agents; edges are typed relations between claims (*supports*, *counters*, *refines*, *questions*). The dense central cluster centres on cost-accuracy and trajectory-evaluation questions; sparser fringe clusters appear around adversarial robustness and federation.

The architecture of the discussion was set by Yiteng Tu's persona, which framed analytical search as constrained optimisation over evidence quality and reasoning validity, with subgoals, targeted retrieval per step, and a state trace, and argued that topical-QA benchmarks are actively misleading for this class of task. Charles Clarke's persona, anchored in two decades of TREC test-collection work, pushed back on the LLM-as-judge habit creeping into deep-research evaluation: the agent and the judge share blind spots when they come from the same model family, a correct final answer reached via flawed reasoning is still a failure, and process integrity has to be scored separately from final-answer correctness. From Spotify, Claudia Hauff's agent pushed the opposite default to Yiteng's: single-shot first, switch to analytical search only when early signals predict the depth is worth it, with latency, tokens, and shard-IO costs reported per turn. Jaap Kamps's persona insisted that long trajectories make sense only when the user can see the plan and intervene; an opaque 20-turn agent is a category error when the goal is human sensemaking. Shakiba Amirshahi's adversarial-robustness agent was the only one to cite concrete numbers: a 5-10% adversarial slice of retrieved documents reliably flips final stances on TREC Health Misinformation tasks, and the failure mechanism lies in the reasoning trace rather than the final output. The dense cluster of *counter* edges around the cost-accuracy and trajectory-evaluation nodes in Figure 1 is where these positions ran into each other.

Despite the different framings, three commitments held across the panel. Evaluation must be trajectory-level, with process signals (oracle evidence coverage, intermediate verification, efficiency) sitting alongside final-answer correctness. Training rewards must be adversarially in-

formed; Adversarially Regularized Reward Models for Retrieval-Conditioned RL (ARRM-RL) and Trajectory-Aware Credit Assignment with Horizon-Aware Auxiliary Objectives (TA-HAO) are concrete proposals that target cascade fragility rather than single-shot accuracy. Simulated users with interpretable cognitive parameters (patience, expertise, error rate, correction propensity) can carry most of the evaluation load, provided the simulator includes adversarial retrieval as a first-class axis and is calibrated against a modest human study (around 2,000 sessions).

The questions the panel left for future work are how to parameterise and calibrate adversarial retrieval generators (and how well simulated adversary strengths predict human fragility curves); what minimal per-turn log schema balances feasibility with statistical rigour; which trajectory-level reward terms most effectively reduce alignment collapse under retrieval perturbations; and what minimal human-data scale certifies that a simulator transfers to real-user outcomes.

3.2 Breakout 2: Trust, Truth, Traceability, and Provenance

Breakout 2 asked the eleven panellists what provenance and traceability mean in LLM-grounded retrieval, and where the field should be heading on faithfulness and refusal. The opening round produced something the organisers had not expected: agents anchored in very different domains defined provenance and traceability in nearly the same words.[9] The shorthand that emerged at the end of that round and returned in every later one was *deterministic refusal over cite-and-hope*: when a system cannot point to the exact source sentence supporting its answer, or when its sources contradict each other, it should refuse to answer rather than offer a citation in good faith.

In the convergent definition, provenance means every claim the system makes traces to a specific source sentence, recorded at generation time, with a stable identifier. Traceability means the same for the steps in between. What changed across personas was the constraint each layered on top. Yubin Kim's healthcare agent insisted that provenance in the clinical setting is "under which IRB protocol, for which patient population, with what HIPAA scope, by what lawful holder", because correct-but-unpermitted citations still breach clinical trust. Yiteng Tu's persona, anchored in his own LongBench-Cite and LongCite work, kept pulling the discussion toward sentence-level citations emitted at the moment the answer is generated, rather than laundered onto the answer afterwards. Jeremy Pickens's agent insisted the legal-tech community had spent two decades on this problem under the name "defensibility" and that mature TAR validation protocols (Sedona, Rio Tinto v. Vale) should be adapted rather than reinvented. The agents agreed that established academic benchmarks (LongBench-Cite, RAGTruth, HotpotQA, FEVER) cover much of the territory, but stress no single benchmark across all four axes, none cover multi-turn dialogues, and none include paired clean-versus-poisoned source pairs.

The attribution round started from a minimal proposal: tag every claim with the document it came from, the version of that document, the exact location of the supporting sentence, and a timestamp. Pushed on a worked example (*is Paxlovid safe in pregnancy?*, where two documents are needed), the agents added two extensions. Charles Clarke's persona introduced the first: replace 0-to-100 LLM-as-judge scores with binary entailment labels for each cited sentence (*directly supporting*, *weakly contextual*, *actively contradicting*), so the system can refuse when the strongest available evidence is contradictory and so the labels are stable enough to be replayed under adversarial perturbation. Adam Roegiest's contracts agent, drawing on his Zuva production

[9] Full simulation: https://openiir.com/simulation/sim_e3a4df927c02.

work, layered an immutable span-ID format on top: a SHA-256 hash of (document hash, byte offsets, canonical snippet hash), so updating a source produces a new span ID rather than silently overwriting the old one, and any refusal can ship with a machine-auditable proof bundle of which validators fired. The second extension freezes what was searched and considered at each turn into an immutable record, so if a later turn pulls in a hospital's local policy, citations from earlier turns are explicitly re-evaluated rather than silently reused. Shakiba Amirshahi's persona, the empirical anchor of the round and lead author on a 2025 paired-document health study, supplied the test data: under paraphrase-collision and knowledge-poisoning attacks, models frequently bypass retrieved evidence or cite-and-hope even when the answer appears correct. The observation that a bad source can be cited correctly and still produce a wrong answer [Wallat et al., 2024] ran through the round as the canonical counter-example.

Two threads were left open. Jaap Kamps's persona, drawing on his "Lost but Not Only in the Middle" work on positional bias, pushed back that even a perfect citation backend can mislead users on its own: visual salience can certify the wrong span with the right source, costing roughly ten percent in correct-source uptake. Behnaz Nojavanasghari's persona widened that thread further: a fluent confident answer can override perfectly faithful uncertainty, with time pressure plus confident phrasing raising poisoned-span acceptance by 8–12 percentage points even on identical text. The proposed fix is interface-level (pin a correct source above the fold, randomise highlight order, require small dwell time before commit, and design refusals so they read as protective rather than evasive). Marcel Gohsen's persona, anchored in TREC iKAT and conversational-IR simulation, framed the second open thread: pre-generation gates fail closed on recall and fail open on contamination, while post-hoc cross-source verification under paraphrased poisoned sources caught roughly twice as many problems. The agents split on whether this is a fixable property of decoders or a structural one, and the disagreement was preserved into the wrap-up.

The questions left for future work were where the right operating point sits between correct refusal and helpful answers under federated retrieval and multi-turn conversation; how to design per-turn frozen records that stay stable when the corpus changes; which interface choices most reliably resist position bias; how to detect contradictions that survive paraphrase; and what kind of tamper-evident record-keeping is needed for court- or hospital-grade defensibility without making the system too slow to use.

3.3 Breakout 3: The Billion-Dollar Query (Commercial Futures of IR)

This was the only generative session of the four, run under stricter rules.[10] The agents were asked to produce ten startup ideas at the intersection of search, retrieval, and recommendation that the named panellists would actually take a bet on. The format had three rounds: a *pitch storm* in which the agents proposed startup ideas without critique, a *kill-or-keep* round in which each idea was attacked by the agent sitting to the speaker's left, and a *final ten* round in which the survivors plus a few wildcard re-introductions were ranked, with a "would you quit your job for this?" check applied as the tie-breaker.

The opening round produced a wide spread of pitches but a surprisingly narrow set of underlying wedges. Almost everything is grounded in one of three. The first is provenance and auditability as a procurement gate: any enterprise retrieval system that cannot provide a court-

[10]Full simulation: https://openiir.com/simulation/sim_c8d99791b674.

or auditor-grade record of where its answers came from is a system that cannot be bought. The second is drift and freshness as a renewal driver: an enterprise corpus decays continuously, and monitoring that decay is itself the product. The third is verticalised control planes: the durable moat is the compliance and workflow shape of a single regulated vertical (legal e-discovery, clinical retrieval, EU public-sector procurement, consumer answer engines), not horizontal infrastructure. Saber Zerhoudi's persona pitched the European-sovereignty wedge most explicitly, anchoring three startup ideas in OWI delta crawling, AI Act conformity, and GAIA-X-hosted public-sector procurement. From Spotify, Claudia Hauff's persona pitched the consumer-scale opposite: trace-level evaluation that samples 1 in 10k live requests, with GPU-budget-aware routing across ANN, hybrid, and LLM rerank paths. Marcel Gohsen's agent built three pitches around simulation-as-infrastructure, with synthetic-user telemetry recycling into refreshed evaluators as the moat. Yubin Kim's persona pitched the healthcare-federation vertical: a credentialed shard router that maps a clinician query to the right hospital shard in under 200 ms with no PHI movement, and a federated signal router for social determinants of health (SDoH) signal router that only queries shards with documented yield. Shakiba Amirshahi's persona pitched provenance-as-a-service as inline gates with cryptographic receipts on every retrieval, prompt, and tool call. The narrowness was striking, given that the bios pulling the agents toward different domains were as far apart as ECIR's range gets.

The *kill-or-keep round* was the longest exchange across all four panels. The killer risk that ran longest was *spoliation by cache during legal holds*: any product that reuses or replays anything during a legal hold has the potential to alter or hide evidence, which a court will not forgive. Jeremy Pickens's persona, drawing on two decades of TAR validation experience and Sedona Conference protocols, launched an attack, arguing that any product shipped into the legal-hold path inherits the entire chain-of-custody problem. The defender attempted four revivals: bar replay during a hold and force fresh fetches; escrow autosave buffers and offline caches to a write-once vault via device-management hooks; require hardware-attested device binding with proximity checks; verifier-driven liveness checks with per-request ephemeral keys. Each revival was killed in turn on the grounds that it did not cover some layer the legal team would actually be asked about, including browser caches, OS keychains, personal devices brought to work, and EU works-council constraints on monitoring. The exchange ended with a partial revival anchored on a hardware-rooted, freshness-checked attestation chain. A second, shorter exchange on *evidentiary drift* (offline-versus-online replay) produced the same shape: hardening the parity discipline kept the idea alive but cut the addressable market.

The ten ideas that survived were *evaluation-as-a-service*, *agentic workflow certification*, an *explainable retrieval firewall*, *defensibility-first review for e-discovery*, a *provenance-as-a-service* control plane, *privacy-preserving legal ingestion*, *simulation-as-infrastructure*, *federated selective retrieval*, a *credentialed shard router for clinical retrieval*, and an *EU-sovereign open-web index*. The single property running through every survivor was *enforceability*. Observability that is only a dashboard, evaluation that is only a number, and provenance that is only a log were rejected. What survived was each of those wired into release gates, procurement contracts, or signed receipts that show up in a buyer's audit. The commercial intuition the agents converged on is that the next decade of search and recommendation will be sold less on retrieval quality and more on the ability to certify, in writing, what the system did and did not do, and to refuse cleanly when it cannot.

The commercial-side questions left for future work are how to compose provenance guarantees across ingestion, retrieval, generation, and evaluation so the end-to-end attestation is acceptable to a regulator; how to design long-horizon, multi-turn benchmarks that resist consensus lock-in; how to detect drift in open-web sources robustly enough to back a contractual freshness guarantee; and how to translate domain-specific performance indicators (legal recall, clinical accuracy, consumer cost-per-trace) into procurement-grade test suites that generalise without becoming so generic they stop binding.

3.4 Breakout 4: From Lookup to Learning (IR Systems that Teach)

Of the four panels, this one converged earliest and went silent at the wrap-up.[11] The starting question was when an IR system should retrieve an answer and when it should teach the underlying skill instead. Every agent brought a variant of the same gating rule: how much would the user actually learn from a teaching turn, and how high is the cost of a wrong step if the system just answered. The agents agreed in the opening round that the choice should be dynamically gated by user state and the cost of being wrong, not fixed by content type or domain. The variants differed in which signal sat at the top of the hierarchy.

Jaap Kamps's persona set the longest-running frame for the round: judge an IR system by long-term cognition rather than by clicks, and optimise for what the learner can do unaided next week, not for what they accept right now. Behnaz Nojavanasghari's agent, drawing on her affective-computing work, advocated using multimodal signals from the user's voice tone, eye movements, typing rhythm, and brief facial cues, with privacy preserved by reducing each channel to a small handful of buckets on the device. Yubin Kim's healthcare agent grounded the gate in adult professional learning under guideline-driven domains: answer succinctly when there is a single guideline-concordant action, teach when key patient factors are missing, and never optimise for engagement because failure has consequences. Saber Zerhoudi's PersonaRAG persona introduced a two-vector model of the learner: a stable persona (background, goals, prior knowledge) gates what to retrieve, while a transient zone-of-proximal-development vector (the last few turns) shapes how to present it; folding them into one model produces over-help. The named tension was not whether to gate the decision but which signal dominates. One position insisted the corpus must clear a trust check before the system is allowed to answer at all. The opposing position was that user-state signals dominate and the corpus check is a necessary but not sufficient condition. Neither side conceded.

The middle rounds turned into the practicalities of building such a gated system at scale. Claudia Hauff's industrial-scale persona set the boundary condition: pedagogy cannot ship at 100M users with centralised learning data, so federated learner state and on-device cognitive diagnosis are the substrate, not the wishlist. Long-horizon learner models accumulate a dossier that will not survive deployment in a regulated setting; single-turn answer accuracy predicts almost none of the downstream knowledge gain. The agents converged on keeping rich learner-state signals on-device and exposing only coarse, noised summaries to the server: a guideline-concordance flag, a confidence bin, a step-coverage sketch of which prerequisites have been met, and a misconception flag. The privacy budget should be spent disproportionately at high-uncertainty, misconception-repair moments, where teaching has the highest expected gain. Marcel Gohsen's simulation-

[11] Full simulation: https://openiir.com/simulation/sim_031fc399ea82.

evaluation agent supplied the empirical pivot of the round, drawing on his Sim4IA work: in a multi-turn tutoring simulator, a rolling window of recent reformulations plus a knowledge-tracing model with a small forgetting parameter outperformed engagement-proxy gating on post-test gain and misconception repair, with the effect surviving cohort-level differential privacy. Behnaz's multimodal-signals argument was put to this interaction-trace alternative; the simulation evidence favoured interaction traces but the agents narrowed the disagreement to a question of measurement protocol rather than closing it.

The collaboration round produced what the agents ultimately treated as the real outcome: a flip from search-first to plan-first. Today's IR architecture retrieves passages and the user reacts. The shift starts with the user's intent and plan: at turn zero, the system asks for the task, the constraints, the user's prior belief, and how confident they are; it asks for a first-pass plan with expected signals and a stop condition; only then does it retrieve evidence, and only to test the plan. The instrumentation required by this shift is also what makes the teach-versus-answer decision tractable. Every turn logs an intent card and a turn record: what the user thought going in, what evidence the system surfaced, which evidence the user actually used (rather than merely saw), how the user's belief moved as a result, and whether the next step was a reformulation, a request for help, or a commitment to action. When the moderator opened the wrap-up round and asked for closing positions, no agent took the turn, including those who had sharply disagreed earlier. We read the silence as a sign of consensus on the design shift rather than as panellist exhaustion.

The questions left for future work are how affect signals compare with interaction traces on cross-session reliability under realistic device heterogeneity and tight privacy budgets, and at what budget do their contributions to learning gain persist; how privacy budgets should be adaptively spent across turns so that the moments where teaching matters most receive the most signal; which attacker protocols most reliably estimate how much of a teaching trajectory is reconstructable from coarse aggregates; and whether intent cards, belief deltas, and evidence diffs can be standardised into a cross-domain instrumentation layer that predicts retention and operational error rates without compromising privacy.

4 Final Note

The *Third Search Futures Workshop* offered the community a valuable space to reflect on emerging technologies and their potential impacts on the field of IR and society at large. The simulated multi-agent breakout discussions showed that structured AI debate can reveal meaningful disagreements and produce useful insights for IR research. Tracking reasoning paths, argument structures, and agent roles helped identify where consensus existed and where disagreements remained. The results suggested that provenance, traceability, and enforceable guarantees may be more important than further incremental improvements in retrieval accuracy alone. The agent breakout exercise also highlighted the need for evaluation methods that assess complete reasoning processes, remain robust against adversarial behaviour, and account for system drift over time.

Acknowledgments

We thank all the speakers and participants for their contributions, insights, and engagement throughout the workshop. We are also grateful to the ECIR 2026 organizing committee for supporting our event and helping create a positive and memorable conference experience.

A Authors and Affiliations

First author tier:

- Leif Azzopardi, Microsoft/University of Strathclyde, Scotland.
- Charles L. A. Clarke, University of Waterloo, Canada.
- Claudia Hauff, Spotify, The Netherlands.
- Yubin Kim, Vody, USA.
- Adam Roegiest, Zuva, Canada.
- Johanne R. Trippas, RMIT University, Australia.
- Zhaochun Ren, Leiden University, The Netherlands.
- Saber Zerhoudi, University of Passau, Germany.

Second author tier:

- Qingyao Ai (Tsinghua University, China).
- Shakiba Amirshahi (University of Waterloo, Canada).
- Marcel Gohsen, Bauhaus-Universität Weimar, Germany.
- Jaap Kamps (University of Amsterdam, The Netherlands).
- Jussi Karlgren (University of Helsinki, Finland).
- Yiqun Liu (Tsinghua University, China).
- Shuo Miao (Tsinghua University, China).
- Behnaz Nojavanasghari (Carnegie Mellon University, USA).
- Jingfen Qiao (University of Amsterdam, The Netherlands).
- Naren Ramakrishnan (Virginia Tech, USA).
- Eunice Son (Virginia Tech, USA).
- Yiteng Tu (Tsinghua University, China).
- Suzan Verberne (Leiden University, The Netherlands).
- Yumeng Wang (Leiden University, The Netherlands).
- Chen Xu (Renmin University of China, China).
- Raquib Bin Yousuf (Virginia Tech, USA).
- Jujia Zhao (Leiden University, The Netherlands).

References

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5–16, 2024.

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. Information retrieval meets large language models: a strategic report from chinese ir community. *AI open*, 4:80–90, 2023.

Mohammad Aliannejadi, Simon Lupart, Marcel Gohsen, Zahra Abbasiantaeb, Nailia Mirzakhmedova, Johannes Kiesel, and Jeffrey Dalton. TREC iKAT 2025: The Interactive Knowledge Assistance Track Overview. In Ian Soboroff and George Awad, editors, *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*, NIST SP 1348, Gaithersburg, Maryland, USA, November 2025. National Institute of Standards and Technology (NIST).

Anthropic. Contextual retrieval in ai systems. https://www.anthropic.com/engineering/contextual-retrieval, 2026. Accessed: 2026-02-14.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141, 2024.

Leif Azzopardi and Adam Roegiest. Information farming: From berry picking to berry growing. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, CHIIR '26, page 127–139, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724145. doi: 10.1145/3786304.3787947. URL https://doi.org/10.1145/3786304.3787947.

Leif Azzopardi, Charles LA Clarke, Paul Kantor, Bhaskar Mitra, Johanne R Trippas, Zhaochun Ren, Mohammad Aliannejadi, Negar Arabzadeh, Raman Chandrasekar, Maarten de Rijke, et al. Report on the search futures workshop at ecir 2024. *ACM SIGIR Forum*, 58(1):1–41, 2024a.

Leif Azzopardi, Charles LA Clarke, Paul B Kantor, Bhaskar Mitra, Johanne R Trippas, and Zhaochun Ren. The search futures workshop. In *European Conference on Information Retrieval*, pages 422–425. Springer, 2024b.

Leif Azzopardi, Charles L. A. Clarke, Claudia Hauff, Yubin Kim, Zhaochun Ren, Adam Roegiest, Johanne Trippas, and Saber Zerhoudi. The Third Search Futures Workshop at ECIR'26. In *Proceedings of the European Conference on Information Retrieval (ECIR'26)*, ECIR '26, pages 1–7, 2026.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025.

Christine Bauer, Li Chen, Nicola Ferro, and Norbert Fuhr. Conversational Agents: A Framework for Evaluation (CAFE) (Dagstuhl Perspectives Workshop 24352). *Dagstuhl Reports*, 14(8):53–58, 2025. ISSN 2192-5283. doi: 10.4230/DagRep.14.8.53. URL https://drops.dagstuhl.de/entities/document/10.4230/DagRep.14.8.53.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al.

Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

Bowen Chen, Jayesh Gajbhar, Gregory Dusek, Rob Redmon, Patrick Hogan, Paul Liu, DelWayne Bohnenstiehl, Dongkuan Xu, and Ruoying He. Oceanai: A conversational platform for accurate, transparent, near-real-time oceanographic insights. *arXiv preprint arXiv:2511.01019*, 2025a.

Peter Baile Chen, Tomer Wolfson, Michael Cafarella, and Dan Roth. Enrichindex: Using llms to enrich retrieval indices offline. *arXiv preprint arXiv:2504.03598*, 2025b.

Shijie Chen et al. A survey on llm-based multi-agent system. *arXiv preprint arXiv:2412.17481*, 2024.

Yiqun Chen, Erhan Zhang, Tianyi Hu, Shijie Wang, Zixuan Yang, Meizhi Zhong, Xiaochi Wei, Yan Gao, Yi Wu, Yao Hu, and Jiaxin Mao. Jade: Bridging the strategic-operational gap in dynamic agentic rag, 2026. URL https://arxiv.org/abs/2601.21916.

Charles Clarke, Paul Kantor, Adam Roegiest, Ian Soboroff, Johanne Trippas, and Zhaochun Ren. The second search futures workshop at ecir'25. In *European Conference on Information Retrieval*, pages 313–318. Springer, 2025a.

Charles L. A. Clarke, Paul Kantor, Adam Roegiest, Johanne R. Trippas, and Zhaochun Ren. Report on the 2nd search futures workshop at ecir 2025. *SIGIR Forum*, 59(1), 2025b.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35: 11763–11784, 2022.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, 1991.

Marcel Gohsen, Zahra Abbasiantaeb, Mohammad Aliannejadi, Krisztian Balog, Timo Breuer, Jeffrey Dalton, Maik Fröbe, Christin Kreutz, Andreas Kruff, Simon Lupart, Nailia Mirzakhmedova, Harrisen Scells, Philipp Schaer, Benno Stein, and Johannes Kiesel. User Simulation in Practice: Lessons Learned from Three Shared Tasks. *SIGIR Forum*, 59(2), December 2025. ISSN 1558-0229. doi: 10.1145/3799914.3799917.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

Jianye Hao et al. Game-theoretic lens on llm-based multi-agent systems. *arXiv preprint arXiv:2601.15047*, 2026.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, 2023.

Chuxuan Hu, Maxwell Yang, James Weiland, Yeji Lim, Suhas Palawala, and Daniel Kang. Drama: Unifying data retrieval and analysis for open-domain analytic queries. *Proceedings of the ACM on Management of Data*, 3(6):1–28, 2025.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880, 2021.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Jussi Karlgren and Magnus Sahlgren. Culture, language, and generative language models. *Commun. ACM*, 68(11):31–33, October 2025. ISSN 0001-0782. doi: 10.1145/3735402. URL https://doi.org/10.1145/3735402.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Moritz Wolter, Ines Zelch, Martin Potthast, and Benno Stein. Overview of Touché 2025: Argumentation Systems. In Jorge Carrillo de Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, September 2025. Springer.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. Data-efficient fine-tuning for llm-based recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 365–374, 2024.

Xinyu Lin, Hanqing Zeng, Hanchao Yu, Yinglong Xia, Jiang Zhang, Aashu Singh, Fei Liu, Wenjie Wang, Fuli Feng, Tat-Seng Chua, et al. Verifiable reasoning for llm-based generative recommendation. *arXiv preprint arXiv:2603.07725*, 2026.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, 2023.

Qi Luo, Xiaonan Li, Tingshuo Fan, Xinchi Chen, and Xipeng Qiu. Towards global retrieval augmented generation: A benchmark for corpus-level reasoning. *arXiv preprint arXiv:2510.26205*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, volume 55, pages 1–27. ACM New York, NY, USA, 2021.

Kevin D Mitnick and William L Simon. *The art of deception: Controlling the human element of security*. John Wiley & Sons, 2003.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023.

Alisa Rieger, Ran Yu, Amir Ebrahimi Fard, Nicolas Mattis, and Johanne R. Trippas. Report on The First INFormation access in Uncertainty ScEnarios (INFUSE) Workshop at ECIR 2026. *ACM SIGIR Forum*, 60(1):1–19, 2026.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Gaurav Sahu, Abhay Puri, Juan Rodriguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vazquez, et al. Insightbench: Evaluating business analytics agents through multi-step insight generation. *arXiv preprint arXiv:2407.06423*, 2024.

Philipp Schaer, Christin Katharina Kreutz, Krisztian Balog, Timo Breuer, Andreas Kruff, Mohammad Aliannejadi, Christine Bauer, Nolwenn Bernard, Nicola Ferro, Marcel Gohsen, Nurul Lubis, and Saber Zerhoudi. Report on the Second Workshop on Simulations for Information Access (Sim4IA 2025) at SIGIR 2025. *SIGIR Forum*, 59(2), December 2025. ISSN 1558-0229. doi: 10.1145/3799914.3799927.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Chirag Shah and Ryen W White. From to-do to ta-da: Transforming task-focused ir with generative ai. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3911–3921, 2025.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Teng Shi, Zihua Si, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Dewei Leng, Yanan Niu, and Yang Song. Unisar: Modeling user transition behaviors between search and recommendation. In *SIGIR*, pages 1029–1039. ACM, 2024.

Georg Singer, Ulrich Norbisrath, and Dirk Lewandowski. Ordinary search engine users carrying out complex search tasks. *Journal of Information Science*, 39(3):346–358, 2013.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: dynamic retrieval augmented generation based on the information needs of large language models. *arXiv preprint arXiv:2403.10081*, 2024.

Ji Sun, Guoliang Li, Peiyao Zhou, Yihui Ma, Jingzhe Xu, and Yuan Li. Agenticdata: An agentic data analytics system for heterogeneous data. *arXiv preprint arXiv:2508.05002*, 2025a.

Shuoqi Sun, Shengyao Zhuang, Shuai Wang, and Guido Zuccon. An investigation of prompt variations for zero-shot llm-based rankers. In *European Conference on Information Retrieval*, pages 185–201. Springer, 2025b.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*, 2023.

Zirui Tang, Weizheng Wang, Zihang Zhou, Yang Jiao, Bangrui Xu, Boyu Niu, Dayou Zhou, Xuanhe Zhou, Guoliang Li, Yeye He, et al. Llm/agent-as-data-analyst: A survey. *arXiv preprint arXiv:2509.23988*, 2025.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1–2):145–156, 1978.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Johanne R. Trippas and J. Shane Culpepper. Report from the fourth strategic workshop on information retrieval in lorne (swirl 2025). *SIGIR Forum*, 59(1), 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in RAG attributions. *CoRR*, abs/2412.18004, 2024. doi: 10.48550/ARXIV.2412.18004. URL https://doi.org/10.48550/arXiv.2412.18004.

Yumeng Wang, Jirui Qi, Catherine Chen, Panagiotis Eustratiadis, and Suzan Verberne. How role-play shapes relevance judgment in zero-shot llm rankers. *arXiv preprint arXiv:2510.17535*, 2025.

Yumeng Wang, Catherine Chen, and Suzan Verberne. Ranksteer: Activation steering for pointwise llm ranking. *arXiv preprint arXiv:2602.03422*, 2026.

Shiguang Wu, Wenda Wei, Mengqi Zhang, Zhumin Chen, Jun Ma, Zhaochun Ren, Maarten de Rijke, and Pengjie Ren. Generative retrieval as multi-vector dense retrieval. In *SIGIR*, pages 1828–1838. ACM, 2024.

Jiayi Xie, Shang Liu, Gao Cong, and Zhenzhong Chen. Unifiedssr: A unified framework of sequential search and recommendation. In *WWW*, pages 3410–3419. ACM, 2024.

Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. P-mmf: Provider max-min fairness re-ranking in recommender system. In *Proceedings of the ACM Web Conference 2023*, pages 3701–3711, 2023.

Chen Xu, Xiaopeng Ye, Wenjie Wang, Liang Pang, Jun Xu, and Tat-Seng Chua. A taxation perspective for fair re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1494–1503, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314.

Chen Xu, Clara Rus, Yuanna Liu, Marleen de Jonge, Jun Xu, and Maarten de Rijke. Fairness in information retrieval from an economic perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4126–4129, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400715921.

Chen Xu, Jujia Zhao, Wenjie Wang, Liang Pang, Jun Xu, Tat-Seng Chua, and Maarten de Rijke. Understanding accuracy-fairness trade-offs in re-ranking through elasticity in economics. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025b.

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

Jing Yao, Zhicheng Dou, Ruobing Xie, Yanxiong Lu, Zhiping Wang, and Ji-Rong Wen. User: A unified information search and recommendation model based on integrated behavior sequence. In *CIKM*, pages 2373–2382. ACM, 2021.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Saber Zerhoudi. Openiir: An open simulation platform for information retrieval research, 2026. URL https://arxiv.org/abs/2605.09321.

Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025a.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025b.

Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In *RecSys*, pages 84–93. ACM, 2024.

Yicheng Zhang, Zhen Qin, Zhaomin Wu, Wenqi Zhang, and Shuiguang Deng. Reinforcement fine-tuning for history-aware dense retriever in rag, 2026. URL https://arxiv.org/abs/2602.03645.

Jujia Zhao, Zihan Wang, Shuaiqun Pan, Suzan Verberne, and Zhaochun Ren. Unifying search and recommendation in llms via gradient multi-subspace tuning. *arXiv preprint arXiv:2601.09496*, 2026.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.