Report on the 2nd Search Futures Workshop at ECIR 2025

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

Paul Kantor
University of Wisconsin Madison
USA
paul.kantor@rutgers.edu

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

Maria Sofia Bucarelli, Xiao Fu, Yixing Fan, Michael Granitzer, David Graus, Maria Heuss, Jaap Kamps, Yibin Lei, Andrew Parry, Damiaan Reijnaers, Maarten de Rijke, Siddharth A.K. Singh, Yubao Tang, Suzan Verberne, Jonas Wallat, Yumeng Wang, Chen Xu, Andrew Yates, Saber Zerhoudi, Jujia Zhao*

Abstract

The Second Search Futures Workshop, in conjunction with the *Forty-seventh European Conference on Information Retrieval (ECIR)* 2025, 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?*

The second 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

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

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: 10 April 2025.

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

1 Introduction

The *Second Search Futures Workshop* [Clarke et al., 2025], held in conjunction with ECIR 2025,[1] 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.[2]

The Second Search Futures Workshop continued the conversation sparked in the inaugural edition [Azzopardi et al., 2024a,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 presented by these technologies and the implications for end-users, systems 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 tone of the workshop was one 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?* In addition, we discussed the major themes discussed in the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025) [Trippas and 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.

[1] https://ecir2025.eu/

[2] https://docs.google.com/presentation/d/1rVFUfX9NW7ehBPs6C7HQRFMLX2hbDn3UF91lgZXfpl8/edit?usp=sharing

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 (see slide deck), talks were grouped around applications, theoretical perspectives, and methodological innovations, each engaging with the opportunities, challenges, and implications for users, society, and IR.

Deep Search: From Matching to Reasoning

Yixing Fan and Maarten de Rijke

With the advancement of information technology and the increasing demand for high-quality information, IR is undergoing a significant transformation from search engines to intelligent information assistants. Conventional search engines rely on keyword-based queries to provide users with relevant documents, whereas today's users demand more personalized, context-aware, and intelligent solutions. Their needs have changed from simply "finding relevant documents" to "solving specific problems" or "supporting complex decision-making". For example, in the healthcare domain, users may require tailored treatment recommendations for specific conditions, while in legal scenarios, they expect concrete action plans or decision-making rationales. This transformation necessitates that IR systems develop a deep understanding of user intent, effectively analyze multi-sourced documents, and deliver customized solutions that truly address users' underlying needs.

However, current search methods still largely adhere to a query-document matching paradigm. They either aggregate matching signals based on lexical comparisons or construct abstract vector representations of queries and documents and compute vector similarity to obtain the final matching score. This matching-based approach falls short when it comes to complex information-seeking tasks. For instance, if we search for "How to maximize a particular investment portfolio?", literal or semantic matching alone is unlikely to yield an appropriate answer. Recently, the continuous enhancement of LLMs, particularly in reasoning, has provided new opportunities and technical support for complex information-seeking. Examples include OpenAI's Deep Research[3] and Grok's Deep Search[4], which enhance the quality of generated results by allowing for extended reasoning periods. This shift from a speed-first to a depth-first approach is steering IR toward what can be termed "slow reasoning."

Meeting these changes requires information systems to develop a deeper understanding of both users' needs as well as documents, moving from relevance matching to causal reasoning. We propose several potential research questions. First, there is reasoning-enhanced intent understanding, which moves beyond semantic expansion to encompass problem planning and decomposition. This approach would allow the system to determine the next search direction based on the current information at hand. Second, reasoning-enhanced content understanding evolves from document ranking to deeper content inference, enabling the system to pinpoint the key elements within documents that truly support answer generation. Finally, logic-guided result generation shifts from ranking documents to producing structured reports. By integrating multi-source information and

[3] https://openai.com/index/introducing-deep-research/

[4] https://grok.com/

distilling core insights into a structured, visual format, this approach aims to significantly improve the efficiency of IR and enhance decision-making capabilities.

Beyond Citations: Ensuring Trustworthy Information Access in the Age of Language Models

Maria Heuss and Jonas Wallat

The way we interact with information is transitioning from direct engagement with source documents towards chat or natural language interfaces that process information into more digestible forms. While these systems promise to make complex information more accessible through simplified language and conversational explanations, their deployment raises significant concerns about trustworthiness, particularly for marginalized communities who have historically been underserved by technological systems and may disproportionately rely on automated advice-giving systems.

A key challenge is the occurrence of hallucinations, where LLMs generate plausible but incorrect or fabricated information, potentially undermining their reliability in high-stakes scenarios [Ahmad et al., 2023]. This necessitates novel ways of ensuring information quality beyond creating trusted databases and safety measures on displayed documents.

While in the early days of LLM-generated chat conversions, users mostly relied on the parametric memory of the model, Retrieval-Augmented Generation (RAG) [Lewis et al., 2020b] has emerged as a solution to heavy LLM hallucinations and a more trustworthy alternative. With a retrieval pipeline as its backbone, RAG promises a grounded way of generating answers with citations that allow users to verify the displayed information. However, citations alone cannot guarantee trustworthiness if they don't faithfully reflect the origin of the generated information.

The field of Explainable AI has long struggled with concerns about unfaithful explanations [Jacovi and Goldberg, 2020] and unwarranted user trust [Rowley and Johnson, 2013], particularly when explanations appear plausible. Users may be inclined to accept intuitively reasonable explanations, even though plausibility does not guarantee faithfulness to the model's actual decision-making process. This parallels challenges in citation verification: while users can theoretically verify cited information, their willingness and practical ability to do so is often limited, especially in complex domains requiring specialized knowledge like medicine and law.

This highlights the need for automated evaluation methods that assess whether cited documents are the proper sources of provided information. In our work, we investigate "post-rationalized citations", where language models generate answers from their knowledge and then search for matching evidence, rather than deriving answers from the documents [Wallat et al., 2024]. Such post-rationalization can be particularly difficult to detect through correctness evaluation alone. We argue that alongside correctness - verifying if cited information can be found within referenced documents - citation faithfulness should be evaluated to determine whether the cited document was actually used during answer generation. This becomes especially crucial in complex domains like legal or medical fields, where even automated correctness measures may struggle.

Moving forward, we need to develop new approaches to assess the trustworthiness of generated information that consider multiple factors: the information source, users' knowledge, social context, interface between user and information, etc., all to ensure a reliable information pipeline that leverages LLM opportunities while avoiding harms. While attention investigation or classi-

cal explainability approaches might provide some initial insight into the reasoning of the model, understanding the actual decision process may require examining model internals through tools like Mechanistic Interpretability [Bereska and Gavves, 2024]. Future work must find such ways to interpret complex language models to ensure safe and reliable information access for all users.

Agentic Information Retrieval: Towards Proactive Knowledge Synthesis

Yibin Lei and Andrew Yates

Traditional IR methods follow the "ten blue links" paradigm, where systems return a ranked list of documents in response to user queries. However, agentic IR [Dalton and Foley, 2018; Zhang et al., 2025] represents a paradigm shift: rather than returning static documents, the system acts as an agent that actively interprets user intent, autonomously gathers information from multiple sources, and integrates it to generate comprehensive answers or even perform tasks.

Leading tech companies have already embraced this shift, introducing products like Grok DeepSearch and OpenAI DeepResearch. Unlike traditional search, these systems do not simply retrieve pages. Instead, they dynamically aggregate and structure knowledge from various sources. For example, when given a query such as "I want to buy shoes in Amsterdam", DeepSearch aggregates contextual information, including real-time availability, store locations, user reviews, and price comparisons, synthesizing them into a structured response, acting as a personalized assistant. While these advancements are largely industry-driven, we argue that academia still holds a crucial role in shaping this field. The long-standing foundations of IR research provide unique domain knowledge that is essential to the development of agentic IR, especially in evaluation methodologies.

The Challenge of Evaluating Agentic IR. Despite rapid industry development, a critical gap remains: How do we evaluate agentic IR systems? Current RAG-based systems still largely rely on text-only, factoid QA datasets like HotpotQA, measured via exact matching. While efforts like TREC-RAG [Pradeep et al., 2024] introduced nugget-based techniques, benchmark queries remain far simpler than real-world information needs. Many require **non-factoid reasoning** (subjective, open-ended, exploratory) and **multi-modal inputs** (text, images, structured data). Effective systems must generate **personalized responses**, adapting to user preferences, history, and evolving intent. Moreover, search is rarely one-shot; **multi-turn interactions** play a crucial role as users refine their queries dynamically. If agentic IR represents the future of search, we must rethink how we define and measure information-seeking success.

A Multi-Agent Simulation Framework for Evaluation. To bridge this gap, we argue for the importance of building multi-agent simulation frameworks to dynamically evaluate agentic IR systems. Instead of relying on static datasets, such approaches can simulate real-world user interactions, enabling adaptive and interactive evaluation. The framework consists of three key components. First, **user simulators**, modeled as intelligent agents with personalized information, including past search logs, preferences, and information about what the user has read in the current session. Instead of static queries, they generate context-aware and evolving interactions,

reflecting realistic user behaviors. Second, **retrieval agents** represent the agentic IR system under evaluation, responsible for proactively retrieving, synthesizing, and presenting knowledge while adapting to user intent. Finally, **information environments** act as external information sources, including databases, web pages, domain-specific repositories, or even domain-specialized LLMs that encode expert knowledge [Lin et al., 2024], mimicking real-world heterogeneous information landscapes.

This multi-agent environment allows for interactive and dynamic evaluation, where success can be measured using techniques like LLM-as-a-Judge [Gu et al., 2025]. Particularly, the framework allows for assessment across four key dimensions. First, **user satisfaction and engagement** evaluates whether the system effectively fulfills complex user needs. Second, **personalization effectiveness** measures how well the system adapts to individual user preferences and past interactions. Third, **task completion rate** determines whether the system enables users to achieve specific goals, such as booking a trip or buying an item. Finally, **information coherence and reliability** ensures that the system-generated responses are accurate, contextually relevant, and, most importantly, trustworthy.

Academic IR research has long been a leader in advancing retrieval methodologies, particularly through its contributions to TREC and other benchmarking efforts. We believe that by leveraging the decades of IR expertise, we can develop more rigorous, scalable, and reliable approaches for both evaluating and improving agentic IR systems.

Let's Use This Wave to Our Advantage, Before it Becomes Our Disadvantage

Damiaan Reijnaers

Recent advances in NLP have led to the wide use of language models in IR contexts [Zhu et al., 2024]. I argue that both their perceived success and failure have created the ideal conditions to now shift focus to building retrieval systems based on more intrinsic, domain-informed knowledge representations. *Why now?*—With 'AI' in the spotlight, domain experts may be drawn to contribute, some inspired by its potential, others prompted by its shortcomings. Their much-needed knowledge may help us build better systems, now, while the momentum is there.

To exemplify my point, I turn to the government context, one in which trust in technology may be especially fragile. The Dutch case, infamous for its "childcare benefits scandal", illustrates both the rapid adoption of algorithm-assisted decision-making and its negative consequences. This, in turn, intensified efforts for government transparency and better public information management. This means that now is both an opportune moment (with more open government data) *and* a necessary one (amid low trust in government and technology) to turn towards more controllable forms of information processing.

Instead of relying solely on textual similarity for case recommendation or search systems, (open) government documents could be represented through more intrinsic features, such as the law articles they refer to, the procedural rules they follow, or the argumentation structures used. Models that extract these features should integrate domain knowledge in their design and could additionally draw on (dynamic) resources, like statutes and government datasets, that are often already available in a structured format.

By leveraging deeper document representations, searching through archival material may become more effective, and document similarity analyses may align more intuitively with legal reasoning. Moreover, the inherent interpretability of the underlying features could help pave the path towards 'explainable similarity' or 'explainable search'. The latter is particularly relevant in the context of the current example, as scholars in public governance are already calling on us to develop algorithms that are *intrinsically* explainable (see, *e.g.*, [Hildebrandt, 2022]). As I already noted, this is not only a point of opportunity, but one of necessity: "Opaque algorithms can undercut people's sense of fairness and trust—particularly when used by the government (...)" [Deeks, 2019, p. 1833].

This, inevitably, brings us back to the fundamentals of AI: the field of study of abstracting reality—much like in mathematics—into an 'artificial', quantitative representation that machines manipulate to achieve practical outcomes in the 'real' world. This position is not about completely discarding the prevailing paradigm; instead, things could go hand-in-hand. The future may lie in hybrid approaches that may bring us into the neuro-symbolic sphere with the development of innovative methods for knowledge representation learning. Such approaches could place greater weight on domain expertise, while still harnessing the power of the past decade's models and doing justice to increasing demands for explainability. But that *does* mean we must go back to the drawing board and learn about how things work in the world around us—and that demands more than a single discipline.

What do You Mean, Information? Semantics, Knowledge Representation, and Natural Language in a World of LLM-powered Search

Siddharth Singh and Andrew Yates

This talk focuses on the potential of text-based neuro-symbolic representations for advancing a number of crucial IR research directions. While some suggest that LLM pretraining inherently lacks a symbolic foundation [Xu et al., 2024c], we challenge this assumption, proposing that natural language itself provides structured opportunities for knowledge representation in IR that have yet to be fully explored.

Building on previous work, we argue that text-based meaning representations provide a computationally efficient path for improving retrieval performance with LLMs [Nie, 2023]. In this context, the vocabulary of an LLM can be viewed as an "upper ontology"—a predefined, abstract space that gives a foundation for describing conceptual relationships in the world. This formulation of the retrieval problem has already found success in Learned Sparse Retrieval (LSR). Approaches like SPLADE [Formal et al., 2021] and Mistral-SPLADE [Lassance and Clinchant, 2023] have been successful using this "upper ontology"; however, only using the vocabulary of the underlying model has shortcomings that can be addressed only by building this bridge. In this case, the bag of words returned by a sparse retrieval system loses crucial morphological information at the syntax-semantic interface due to the nature of LLM tokenizers. Moreover, this sparse representation has a difficult time accounting for synonymy and homonymy, requiring an approach to disambiguate between different word senses.

One promising text-based, neurosymbolic approach to representation is the use of lexically-grounded, semantically-driven computational resources that contain syntactic or conceptual information. A viable example is FrameNet [Baker et al., 2001]. Unlike its predecessor WordNet

[Miller, 1998], a lexical database that only provides a vocabulary/thesaurus function, FrameNet is a lexical database constructed on a semantic theory which defines words based on "scenes" where the word appears and the thematic roles it assumes in the scene; "buy", "sell", and "goods" are all parts of a semantic scenario called "commercial transaction".

This approach both disambiguates word senses and re-adds the syntactic directionality removed by a sparse bag-of-words representation. It is also well-equipped to handle linguistic negation, a major challenge in IR for both sparse and dense representations [Weller et al., 2024]. This issue thus presents an opportunity to explore multi-model solutions to the negation problem: for example, a specialized "negation expert" model could work in tandem with general retrieval models, ensuring that certain terms, roles, or relationships that should not appear in a retrieved document are correctly identified and weighted. In the same vein, this and other text-based neuro-symbolic representations are robust to both changes in language and changes in model architecture; queries and documents can be encoded by different models so long as the symbolic system remains the same.

The discussion will then focus on two main research directions: (1) developing effective text-based representations by effectively combining ontologies, knowledge graphs, and lexicons like the aforementioned resources, and (2) designing specialized datasets and models that use these text-based neuro-symbolic representations, specifically in an IR context. Current limitations in IR systems highlight the need for better neuro-symbolic representations that capture relevant signals/meaning across both languages and domains. Ultimately, we stress the advantage of efficiency, robustness, and interpretability as advantages of these approaches.

Beyond Text: Integrating Multimodal Information into Generative Retrieval

Yubao Tang and Maarten de Rijke

Generative retrieval (GR) has emerged as a promising paradigm that unifies indexing and retrieval within a single generative model [Metzler et al., 2021; Tay et al., 2022; Tang et al., 2024b]. In this framework, corpus knowledge is encoded directly into model parameters during training, enabling the model to generate relevant document identifiers autoregressively given a query. This mechanism simulates a form of learned memory, eliminating the need for explicit external indexing. GR has demonstrated strong performance in text-only retrieval tasks [Chen et al., 2022; Zeng et al., 2024; Tang et al., 2025, 2024a], especially under low-resource and end-to-end learning scenarios.

However, real-world information needs are often inherently multimodal, involving both textual and visual content [Liao et al., 2018; Cao et al., 2020]. To support such needs, it is critical to extend GR into the multimodal domain—resulting in what we term generative multimodal retrieval (GMR). GMR aims to unify generative indexing and retrieval across modalities within a single architecture. This extension, while conceptually appealing, introduces several unique challenges: the model must not only encode and retrieve across heterogeneous data types (e.g., images and text), but also maintain coherent semantic alignment between modalities.

Some recent efforts have explored leveraging generative models for tasks involving both text and images, but these approaches are often designed for specific applications, such as question answering or document retrieval conditioned on multimodal queries [Long et al., 2025]. In contrast, the focus of GMR is on a more general multimodal retrieval setting, where a unified generative

model directly retrieves both document and image identifiers given a query. Unlike previous methods that focus solely on textual retrieval or answer generation.

The development of an effective GMR framework necessitates careful consideration of several key research questions: (1) Identifier representation. How should identifiers be designed to capture modality-specific and cross-modal semantics? Should image and text representations from the same document share a unified identifier, or should they be assigned distinct codes? (2) Model architecture. What structure best supports multimodal generation? Should the model adopt a decoder-only or encoder-decoder architecture? Should modality-specific decoders be used, or is a single shared decoder sufficient? (3) Training strategy. How should the model be trained to effectively index multimodal content? Should modality-specific representations be learned independently, or jointly, to encourage semantic alignment and fusion? and (4) Inference strategy. Given the identifier structure and generative nature of GMR, how can constrained decoding be employed to ensure both efficiency and correctness in retrieval [De Cao et al., 2021]?

Addressing these questions lays the foundation for a new class of retrieval systems that are both generative and multimodal. GMR has broad potential in real-world applications such as product search, multimedia recommendation, and digital libraries—where retrieving and understanding both textual and visual content is crucial.

This perspective highlights the need to bridge generative retrieval with multimodal understanding and points to promising research opportunities at their intersection. As large language and vision models advance, scalable and robust GMR systems will be central to the next generation of search technologies.

Enhancing User Trust in Conversational IR: The Role of Explainability in the Age of LLMs

Yumeng Wang and Suzan Verberne

In the age of LLMs, conversational IR is becoming increasingly sophisticated, enabling systems to generate human-like responses and provide users with dynamic, engaging search experiences. Yet, challenges such as hallucination, information overload, and the need for transparency highlight the importance of explainability for building trustworthy systems [Huang et al., 2023; Anand et al., 2023].

Explainability is critical in conversational IR [Mo et al., 2023, 2024] for several reasons. First, from the knowledge source perspective, LLMs often integrate external databases, such as RAG [Lewis et al., 2020a], to provide up-to-date and accurate answers. Proving the reliability of these sources and presenting them to users is essential for fostering trust. Second, from the model perspective, LLMs are prone to generating plausible but false information due to hallucination. Explainability mitigates this issue by revealing the model's reasoning process and verifying whether outputs align with the underlying data. Third, from the user perspective, conversational IR involves dynamic, multi-turn interactions where user intents evolve. Success depends on the system's ability to align with these shifting intents. Explainability allows users to track how their queries are interpreted and ensures that the system adapts to their evolving needs, enabling them to adjust queries and achieve their goals effectively.

A central question is what to explain to users. While many studies focus on developer-facing explanations, user-facing transparency is equally important. This raises the question of what users

value most and what information should be displayed. A broad classification of explainability in conversational IR can be: (1) From system output to system input [Qi et al., 2024]: Attributing where the information comes from, such as linking to external sources (e.g., RAG); (2) From user input to system input [Wang et al., 2024b]: Explaining how the system interprets user queries, including intent understanding and query reformulation; and (3) From previous system output to current system output: Highlighting which historical turns are relevant to the current query, how intents shift, and how these shifts impact the conversation's final goal.

However, implementing explainability in conversational IR is not without challenges. First, different stakeholders value different aspects of explainability, making it a highly subjective endeavor. Second, the importance of explainable components varies depending on the conversation's purpose. Complex queries with contextual information require a higher level of alignment between the user's intent and the system's understanding, which is more challenging to achieve and explain. In contrast, exploratory searches with simple keyword-based queries may require less complex explanations. Balancing these varying needs while maintaining system efficiency and usability remains a significant challenge.

In conclusion, explainability is a cornerstone of trustworthy conversational IR systems in the age of LLMs. By addressing the challenges of hallucination, information overload, and user intent alignment, explainability not only enhances system reliability but also fosters user trust. Future directions can focus on tailoring explainability to user needs, addressing subjective stakeholder preferences, and adapting to the diverse requirements of different conversational contexts. In the near future, we aim to explore these dimensions and propose actionable solutions for integrating explainability into conversational IR systems effectively.

How to Balance the Needs of Stakeholders in IR Ecosystems

Chen Xu and Maarten de Rijke

In IR ecosystems, multiple stakeholders typically coexist, including users, content providers, advertisers, and platforms [Abdollahpouri et al., 2020; Abdollahpouri and Burke, 2019]. Different stakeholders in IR systems play distinct roles: (1) users provide feedback on the presented ranking lists, (2) content providers supply documents or products for the platform, (3) advertisers aim to maximize the visibility of their advertisements, and (4) platforms seek to generate profit within this ecosystem. IR systems should ensure that all stakeholders involved can benefit and establish incentive mechanisms that motivate each party to contribute to maximizing collective gains, ultimately achieving a win-win situation for all parties. To achieve this, we believe that IR research should further focus on three key areas of investigation.

1. **Multiple objective optimization**. Given the limited ranking slot and user attention are often limited, it is often challenging for an IR system to fully meet everyone's needs. Therefore, ensuring a fair distribution of IR resources among all stakeholders is essential. For instance, we address the balance between users and providers, commonly referred to as the provider fairness problem [Xu et al., 2023, 2024a, 2025], as well as the balance between users and advertisers [Feng et al., 2007]. However, effectively formulating the interests of different stakeholders remains a significant challenge. It is still unclear how to accurately model and balance these potentially conflicting objectives in a principled manner. Moreover, the key factors or desirable properties that should be considered when designing the

optimization objective, such as continuous, controllable [Xu et al., 2024a], robustness [Deb and Gupta, 2006], or interpretability [Wang et al., 2024a], have not yet been fully explored or theoretically grounded.

2. **Long-term evaluation.** The performance of these objectives should be evaluated in dynamic, long-term IR interactions, where the impact on stakeholders evolves over time. For example, the evaluation should consider the feedback loops [Xu et al., 2024b], which refer to the cyclical interactions between users and the system, where the system's outputs (e.g., rankings or recommendations) influence user behavior, and user behavior in turn affects future system decisions through logged interactions. How to accurately model the long-term interaction process between users and IR systems remains a largely unexplored problem.
3. **Ecosystem Simulation.** In order to explore how to evaluate the long-term performance of the aforementioned complex IR ecosystem systems, we need to build a simulator for IR. Among them, a practical way is to utilize LLMs-based agents [Zhang et al., 2025]. Existing simulators struggle to model how IR systems evolve under the influence of different stakeholders. The era of LLMs presents an opportunity to conduct accurate and fair evaluations of the optimization objectives mentioned above. However, using agentic IR as a simulator faces key challenges such as the difficulty of modeling complex and often irrational user behavior, the presence of multiple interacting stakeholders et al.

In conclusion, to effectively balance the diverse needs of stakeholders within the IR ecosystem, we highlight three essential directions: defining tailored optimization objectives, conducting long-term evaluations, and developing powerful ecosystem simulators. These components are complementary and collectively play a crucial role in protecting and aligning the interests of all involved stakeholders.

Serendipity Engines: Exploring Proactive Web Search via LLM Agents, RAG, and Simulated User Feedback

Saber Zerhoudi and Michael Granitzer

The traditional paradigm of search engines has long centered around explicit user queries—the "ten blue links" approach requires users to articulate their information needs clearly and precisely. However, in real-world scenarios, information needs often emerge organically through user activities, sometimes before users themselves recognize these needs explicitly. We argue that the integration of LLMs offers an opportunity to fundamentally re-imagine this paradigm through what we term "Serendipity Engines"—context-aware systems capable of proactively suggesting relevant information based on implicit user feedback.

While complete anticipation remains highly challenging, a more tractable approach involves leveraging LLMs to create web search experiences that are context-aware and proactively suggestive, rather than relying solely on reactive, query-based interactions [Azzopardi et al., 2024a]. If traditional search engines are limited by the user's ability to articulate their needs explicitly, "Serendipity Engines" should leverage implicit user feedback to provide relevant information and suggestions without requiring explicit queries.

Unlike purely predictive systems that attempt to anticipate user needs without context, we propose leveraging observable user activities within a computing environment as signals for po-

tential information needs. With appropriate user consent and privacy safeguards, implicit signals such as application usage patterns, document interactions, keystrokes, and browsing behavior can provide rich contextual information about a user's current task. These signals, when processed through LLM-agentic systems, can generate contextually relevant suggestions.

Frameworks like SearchLab [Zerhoudi and Granitzer, 2025] offer promising platforms to investigate these systems. Potential experimental designs could incorporate: (1) an LLM-agentic system functioning as a personal search agent that analyzes implicit user feedback to infer context and potential tasks; (2) RAG techniques synthesizing information from retrieved sources while maintaining attribution; and (3) user simulation approaches to generate realistic user profiles with varying research behaviors and cognitive styles.

The technical implementation of such systems presents several research challenges. First, the LLM must effectively interpret diverse implicit signals and translate them into potential information needs. Second, retrieval mechanisms must balance relevance with discovery, preventing overspecialization that might reinforce existing knowledge at the expense of novel insights. Finally, the presentation of suggestions must be minimally disruptive while remaining accessible.

"Serendipity Engines" represent a potential evolution in how we conceptualize search—not merely as a tool for answering explicit questions, but as an intelligent companion that enriches information interactions through contextual awareness. By focusing on this middle ground between purely reactive and purely predictive systems, we may discover new approaches to information access that better align with how humans naturally work and learn.

Ultimately, the success of these systems will depend not only on their technical capabilities but on their ability to integrate seamlessly into users' information ecosystems while respecting privacy boundaries and maintaining user agency.

Unify Search and Recommendation in the Generative Era

Jujia Zhao, Zhaochun Ren, and Suzan Verberne

Recommender systems and search engines have become indispensable components of modern online service platforms, including e-commerce websites and social media networks [Zhang et al., 2024; Wu et al., 2024]. Traditionally, search and recommendation (S&R) tasks are trained using separate models [Bhattacharya et al., 2024]. However, unifying these two tasks within a shared model presents significant advantages [Yao et al., 2021]. First, jointly modeling user behaviors across both S&R provides a more comprehensive understanding of user preferences. Second, leveraging interactions from both task domains enriches item-side representations, improving item understanding and delivery. Therefore, developing a unified model for S&R is a promising direction.

The emergence of generative models, particularly LLMs, introduces transformative opportunities for unified S&R modeling. Leveraging their advanced reasoning capabilities and sophisticated contextual understanding, LLMs can interpret complex user queries and diverse behavioral signals more effectively. Consequently, they enhance the quality of both user and item embeddings, driving improvements in performance across S&R tasks.

Future directions for unifying S&R include the following:

1. **Item identifier design.** Developing item identifiers that encapsulate sufficient information suitable for both S&R tasks is crucial. In search, item representations depend primarily on semantic matching between queries and items, as users explicitly express their information

needs. In contrast, recommendation relies more on collaborative filtering (CF) signals derived from historical interactions, as users passively receive item suggestions without clearly defined intent. Therefore, it is crucial to incorporate both semantic and collaborative signals into item identifiers to support both tasks effectively.

2. **User behavior focus.** Differentiating and balancing user behavior emphases between S&R tasks is essential. Search tasks prioritize immediate user intent captured in short-term behavioral contexts, while recommendation tasks require the integration of both long-term stable preferences and recent interactions. Developing methods to dynamically adjust behavior modeling according to task-specific contexts will significantly enhance the adaptability and effectiveness of unified models.
3. **Efficiency optimization.** Improving model performance without escalating computational costs remains a critical challenge. Unified models inherently process longer user interaction sequences due to combined task data. Therefore, it is crucial to efficiently select and weight relevant user behaviors based on query context, loss indicators, or other discriminative signals to ensure both computational feasibility and predictive accuracy.

In summary, integrating S&R into a unified generative modeling framework, particularly utilizing LLMs, holds immense potential to enrich user and item understanding and improve online service quality. Future research should focus on innovative item identifier designs, task-aware user behavior modeling, and efficiency optimization strategies to fully harness the advantages of unified S&R modeling.

3 Breakout Group Summary

3.1 Semiotics: IR and Meaning

3.1.1 Linguistics

Recent thinking in natural language processing has increasingly moved toward understanding retrieval as a problem rooted in syntax and semantics. Rather than treating queries and documents as bags of words or isolated vectors, this perspective considers how structural and interpretive layers of language affect meaning and relevance. In particular, integrating ideas from interactional sociolinguistics can improve conversational search systems by better modeling the social and pragmatic dimensions of dialogue. These systems could be more attuned to how meaning shifts depending on the speaker, context, and communicative intent. Additionally, current models often miss subtle semantic features, such as connotation, implicature, or emotional tone—dimensions of meaning that strongly shape interpretation but are rarely captured in current embedding spaces.

3.1.2 Anthropology and Linguistic Anthropology

From an anthropological perspective, representations of language in information systems could benefit from a deeper understanding of how signs and meanings vary across cultures and social groups. Concepts from semiotics and linguistic anthropology highlight that the same phrase or symbol can have vastly different interpretations depending on cultural background or group membership. Incorporating these insights could lead to retrieval systems that are more sensitive

to variational sociolinguistics—how language differs across communities in terms of vocabulary, pronunciation, and discourse patterns. This also suggests the value of using discourse analysis not just as an analytical tool, but as a design perspective: one that attends to how users construct meaning in context and in interaction.

3.1.3 Psychology

Information systems typically model users as rational agents with static preferences. A psychological perspective, however, emphasizes how meaning and language use vary across individuals and developmental experiences. Factors like personality, disposition, and upbringing can significantly shape how people express themselves and interpret information. For example, users from similar family environments might exhibit distinctive communicative styles or linguistic habits ("don't point at me!"). Capturing these subtle differences could allow for more personalized and context-aware retrieval. Computational stylometry offers a technical pathway into this space, especially in modeling idiolects—the unique linguistic fingerprint of an individual. These patterns might be used to fine-tune systems not just for demographics or clusters, but for individual users, treating each person's language use as a key to their information needs.

3.2 Agentic IR

Agents in IR systems are still relatively new in function and role, with ever-increasing research and development going into figuring out how best to use them. One of the most poignant examples of agentic IR are Google's DeepResearch[5] and NotebookLM[6] systems that allow users to either offload research "entirely" or work collaboratively with the system. But these agents still rely entirely on human direction, intervention, and often review to ensure that the work they do actually results in the desired outcome.

As the Search Futures workshop seeks to look further into the future than just tomorrow or next week or next month, we discussed the possibilities that would happen when agents are truly capable of having one or more objectives (potentially of their own making), the ability to learn and use new tools, and are autonomous, independent, can function asynchronously, and are capable of reflection both on any plans they made but also in light of human feedback. In this future, we expect that agentic systems will be able to model individual user preferences at a sufficiently high level of fidelity. This level of fidelity is likely to be required to do more complex information processing tasks (e.g., creating different itineraries capturing different user interests and requirements).

In some ways, the agents of this (potentially not so far) future can become proxies for individuals themselves and may allow seamless task completion with direct human intervention. For example, two humans agree to go see a particular musical band, and their agents work together to figure out the best date, the optimal seating, travel arrangements, and so on, without necessarily requiring direct prompting by humans. It is immediately apparent how useful such agents could be in our day-to-day lives, but there are negatives to this outcome, which warrant further discussion.

[5] https://gemini.google/overview/deep-research/
[6] https://notebooklm.google/

The remainder of this section is broken down into discussing the gap between the existing state of agents and user simulation and that which is needed for these powerful agents to be possible, the dangers in customizable or trainable agents, and the dangers to human self-actualization.

3.2.1 Bridging the Gap

Current user simulation approaches, while foundational for developing agentic IR systems, fall short of what is needed for truly autonomous agents. The central gap appears to be in the complexity and fidelity of user modeling. Existing agent simulations [Maxwell and Azzopardi, 2016; Zerhoudi et al., 2022b; Wang et al., 2025] face critical gaps in modeling human-like decision making. First, they rely on **oversimplified preference models** that fail to capture nuanced or contradictory human choices, neglecting deeper motivations, values, and contextual influences. Second, these agentic systems lack **dynamic adaptation mechanisms** to evolve alongside users, despite human preferences being refined through experience and shaped by latent factors (e.g., environmental shifts, social interactions). Finally, current agents focus more on individual users while ignoring the **broader social and environmental contexts** that users live in and interact with (e.g., cultural norms, peer actions, resource constraints) that fundamentally shape their decisions.

To realize the potential of agents, we will likely need to revisit multiple research domains. We suspect that more nuanced means of capturing implicit and explicit signals of user preference will need to be developed (e.g., aligning facial expressions to IR task outcomes). We also need to improve computational modeling to better represent the hierarchical and contextual nature of human preferences (e.g., understanding that preferences in one domain may contradict those in another without being inherently inconsistent). Finally, we need to improve learning methodologies so that agents are enabled to incorporate new insights into their users' preferences over time without catastrophic forgetting of important insights.

Measuring the success of our "futuristic" agents is a challenge on its own. Traditional metrics [Kelly, 2009; Harman, 2011; Zerhoudi et al., 2022a] might be insufficient. Instead, there is likely to be a need for new evaluation frameworks that assess how closely an agent's decisions align with what the human would have chosen given perfect information and unlimited cognitive resources (or approximations thereof). Perhaps most importantly, measuring how well users maintain their sense of agency and control while benefiting from agent assistance will be critical in addressing concerns such as self-actualization.

3.2.2 Dangers in Agent Customization

One of the discussed avenues where agent customization could yield less than desirable outcomes is in the context of purchasable "agent personas." Essentially, there would potentially be a market for agents that resemble popular figures or other scenarios (e.g., life coach-based advice). While there is some potential entertainment value for these personas (e.g., "how would my favorite celebrity go on vacation"), there is a risk that in doing so, we may see further intensification of echo chamber effects, especially if these personas are made by less than scrupulous individuals. Conversely, they may help individuals to make decisions that help them achieve goals (e.g., healthier lifestyle choices by using an agent designed by Arnold Schwarzenegger).

If agents are allowed to interact with the wider world without direct human guidance and are able to "modify" themselves (e.g., in current terms, rewriting their system prompt), then we can envision situations where this can run afoul. The obvious one is that agents, themselves, become targets for influence (e.g., echo chambers, mis-/disinformation) and that this can unknowingly propagate to their human user. Similarly, there is a need to ensure that appropriate controls are in place to ensure that agents do not fall prey to social engineering attacks (e.g., phishing) that might reveal information a user does not want revealed.

Agents interacting in networks might establish sophisticated communication protocols that optimize for efficiency rather than human interpretability. These could include utilizing encoding methods that hide information from human oversight without malicious intent, simply through optimization processes where certain communication patterns yield better outcomes. Indeed, there has been evidence that different language patterns can emerge when models are left to their own devices.[7] Such optimizations may then limit how much sense humans can make in overseeing these interactions and result in a lack of trust in the results.

Several issues highlighted are dependent on how much information is provided to one's agent and how dissemination of that information is controlled. We might reasonably imagine that specifying trusted agents (or organizations) might be part of these agents to ensure that a user's information is not disclosed to those that they do not wish it to be. The risk is that this may limit the quality of an agent's results and outputs to their user, and so helping users determine appropriate levels of the information release is likely a prerequisite to successful adoption of these models, especially in light of humans being less rigorous about digital safety than might be advisable.

3.2.3 Dangers to Self-Actualization

There are obvious benefits to agents that are able to interact in various capacities for their human users. Whether this is from finding and perhaps preparing bureaucratic documents (e.g., taxes) to helping users plan trips or purchasing decisions, to proactively providing information (e.g., restaurant recommendations for travelers at an appropriate dinner time). The danger to this is that users become complacent in their decision making or overly reliant on their agent, and that when failures happen that users are taken by surprise (e.g., fines for late filing of taxes).

Related to this is a potential loss of control or decision-making to the user. If these agents can model the user's preferences to a reasonably high degree of fidelity, does the agent need to actually involve the user? In our discussions, we found that maintaining control of an agent's behavior is a critical component to ensure that the user isn't spoonfed results or answers, but can help inform them.

Even then, there remains a risk that these decision points allow a user too much self-reflection, which may cause emotional or cognitive harm. For example, an eco-conscious user might find their agent repeatedly suggesting products or activities they enjoy-but that harm the environment-based on their preferences. Arguably, a smart agent may even "game" the user by offering environmentally friendly options in conjunction with the more likely to be selected but environmentally un-

[7] As seen in Facebook's negotiation models. From https://www.independent.co.uk/life-style/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html. Accessed May 16, 2025.

friendly option. In doing so, the user is given an illusion of choice that reinforces their self-image but ultimately runs counter to the user's underlying motivations.

We note that this is not necessarily negative. Such an agent could also be used to reinforce a user's desires as well (e.g., only providing environmental options despite a "better" option being available). While this does result to some degree in a potential loss of self-determination by the user, this may not be an outcome that a user would ultimately be opposed to, depending on how this manifests (e.g., if it helps them meet a health goal).

To effectively and appropriately influence users, it is crucial for agents to determine both the type and level of detail of the information they provide. This involves a trade-off between computational cost and the credibility of the information: more detailed content should increase user trust and reinforce their existing beliefs. Additionally, different sources of information (e.g., scientific papers, online forums, social media) will influence users to varying degrees depending on each users' view of those sources. This means that agents may have to account for latent preferences that they themselves may not be aware of (e.g., echo chamber effects).

These issues raise an important focus on how objective functions for agentic behavior are designed as they can impact user control and autonomy. While agents may maximize user satisfaction, they might also create an illusion of control rather than actual empowerment. The tension between optimizing for stated preferences versus revealed preferences might open up the door for a new research direction: Should an agent follow explicit instructions or adapt behaviors that contradict those instructions if it will make its human happier?

3.2.4 Concluding Remarks

In our discussion, we continually returned to ensuring users have the ultimate control over how their agents behave, regardless of how many agents there may be and how autonomous those agents may be. By ensuring that control is obvious and explicit then users can provide their own goals and desires to their agents as needed. Such an approach may also make it clearer to users when things drift from their goals and may aid them in correcting that behavior. While this may not work for all users (e.g., laissez-faire users just want stuff done for them), we find that it provides a reasonable user experience across all users.

A more extreme solution that was discussed is for agents to be separated into domain-specific agents that coordinate through the user explicitly rather than asynchronously, without user awareness (or consent) as proxies. This fragmentation into purpose-built, domain-specific agents could develop deeper expertise within a specific domain while maintaining clear boundaries around data usage and permissions. Users might find it easier to understand and control agents with narrowly defined purposes rather than grasping the decision-making of a complex agent handling all aspects of their digital information. This approach maps naturally to existing mental models of delegation, where humans typically assign different tasks to different experts. The coordination between these specialized agents could occur through explicit user mediation, which reduces the risks associated with autonomous inter-agent information sharing.

3.3 Stuff's Getting weIRd

Gathering around a table in a nearby café, the weIRd group set themselves no limits in thinking about what the future might bring. Nothing was too wild or crazy to consider. Quite the opposite:

Ideas were dismissed for being too obvious or realistic. We clustered our thoughts into three themes, each taken to the extreme: (1) extreme interfaces, (2) extreme personalization, and (3) extreme information needs.

3.3.1 Extreme Interfaces

We asked ourselves: What is the ultimate interface for an information access system? Clearly, it is not typing on a keyboard or even talking to a device. Unless, that is, the device is **fully embodied**, able to interact with the physical world, as well as with the information space.

We imagined a floating head, hovering next to you as you go about your life. If you have a question, you ask it, like a friend. You can send it off to do tasks, both in the physical world and in the information space. What is on the menu at that restaurant? It looks up the answer in the information space, but also floats off to ask a waitress (possibly another AI embodiment): What is the special of the day? It speaks Italian, even if you don't. Fly up and take a picture of us in this piazza. Post it on Instagram. Tag it with our names.

Conversation with your floating head should be **fluent and effortless**. Conversing with current information access systems is robotic, requiring a wake word ("Hey floating head") and following a strict command/response structure. At the same time, current systems can misunderstand commands, causing the user to enter "bad dog" mode, where the command is repeated with increasing volume and vehemence ("Order another beer. Another beer. Beer. Another beer.") Talking to your floating head should be like talking to your very close friend. It just gets you, completing your thoughts before you finish speaking, and even interrupting you to point out interesting sites or sights, or to warn you about a potentially dangerous action.

At the extremest extreme, perhaps we don't converse at all. Perhaps the ultimate information access system is a **direct neural link**: You think what you want, and the answer appears in your mind. Your floating head is no longer a separate embodiment, but a second embodiment of you. You want to see from above the piazza or what is on the menu of that restaurant, you see through the eyes of your floating head. You and the information are one.

3.3.2 Extreme Personalization

Whether embodied or not, future information access should be fully personalized. Ultimately, we each will have a language model that's fine-tuned to us, a full **digital twin**. Recommendations are made by asking the twin. You don't need to be consulted. No need to read a menu or order another beer – the restaurant's AI will ask the twin, and it will appear. The information access system will know what you don't know, will know what you want to know, and will know what you don't want to know. Information will be tailored to you, and just for you, fully reflecting your current knowledge and interests.

Unfortunately, extreme personalization creates risks. Since your twin knows everything about you, reading every email, listening to every conversation, it must **protect your privacy** in the way that you yourself would. The bartender doesn't need your life story along with your beer order. Without care, a digital twin may also place you in a filter bubble from which you can't escape, a **filter jail**. If it knows what you want to know, and tells you only what you want to know, then your view of the world may never change. Every bias will be confirmed, every assumption reinforced. You are always in your happy place.

3.3.3 Extreme Information Needs

Future information access should be as anticipatory and **pre-emptive** as possible. It doesn't just know the answer before you ask the question, it should know the next question, and the answer to it, as well. It knows when to answer questions that are unasked. Many years ago, the attendees at SWIRL 2012 [Allan et al., 2012] imagined a world with "someone's phone ringing as they walk down the street, interrupting their thoughts with the message that the love of their life is sitting in café they are just walking past. In this case, the urgency of the information need is judged to outweigh the annoyance of the interruption. In order to reach this level of performance, deep insights into personality and preference are required."

As this imagined world grows closer to reality, future information access must allow queries that go beyond the merely factual to encompass **physical, spiritual, and emotional** components, ranging from "Where's the nearest toilet" (based on the physically detected query caused by too much beer) to "How do I find true love?" (perhaps stemming from the same cause, or perhaps reflecting a deeper need). Queries (spoken, unspoken, or anticipated) might include spiritual components, with the system becoming almost a digital priest ("Go to this church in this piazza. It will give you the solace you require.")

Answering extreme information needs also creates risks. If your beer order reflects your physical state, perhaps your digital twin or floating head will order a non-alcoholic beer — and not tell you. If your request for pictures of your trip to Italy includes pictures of your former love, perhaps the system will omit them. Slowly, the painful parts of your life will fade away. Your information access system becomes your **caretaker**. Your filter jail will encompass your entire life. You are always in your happy place, with no way to escape.

3.4 Pre-emptive IR

Having to travel home early, a group of three participants formed and continued their discussion on the bus to the airport. Loose brainstorming led to the topic of "pre-emptive IR" and the aspects of a responsible implementation of it.

In Figure 1, we give an overview of two dimensions that contribute to the way that users receive (or might not currently receive) information:

- Whether or not the user is aware of the existence of a piece of information ("information need").
- Whether or not the user wants the information (or would want if they knew it existed) ("information awareness").

Interestingly, these dimensions express the interplay between *user awareness* (of search intent) and *system awareness* (of a user's interest to consume information), with the overlap itself subject to other dimensionalities, such as *relevance*. We identify 4 quadrants that contain different kinds of information that we will discuss below.

1. **The user has an active information need (and is aware of the information).** The top-right quadrant in Figure 1 represents this traditional information-seeking scenario; here,

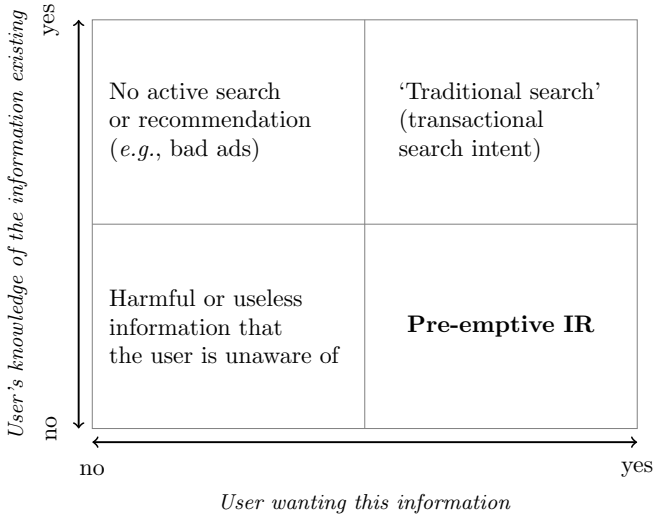

Figure 1. Grid of user need and availability awareness.

the user has a clear information need that can be translated into a search intent. Classic or conversational search can be used to satisfy the user's need.

2. **The user is aware of information but does not want it.** In the top left, we present the scenario where even though the user is aware of the existence of the information in a more or less abstract way, they are not interested in receiving it. Scenarios where the user might be presented with information from this quadrant could be incorrectly retrieved or recommended documents, but also consider targeted bad ads (such as repeated ads, or ads that have lost their relevance), or propaganda that the user would rather not be shown.
3. **The user is unaware of the existence of the information, but also does not want the information.** This quadrant (bottom-left) concerns information the user might know about, but also not want to even know about the existence. Examples of this include triggers you didn't know would have triggered you, 'bad news' at inconvenient times (*e.g.*, a paper rejection), or information hazards. This list may be extended with many more examples that we are unknowingly happily unaware of. Presenting the user with this kind of information at all (or at the wrong time) may be annoying or might even be harmful.
4. **The user would want to have the information but does now know it exists: Pre-emptive IR.** In this scenario (situated in the bottom-right), a user might be interested in receiving this type of information, but their unawareness of its existence makes it impossible for them to actively seek it. We argue that this is a new but promising scenario for IR applications with already existing examples, such as: (1) push messages on the phone about relevant sights in the area, (2) a close and personal friend, who knows your interests and shares interesting information of any kind, unbounded by a specific information type or underlying collection (e.g., book titles, news events, recipes, sights), perhaps closest to the latter, and (3) "good" online ads (i.e., a surprising ad for a product or service you were unaware of, but are interested in).

While seemingly overlapping with a common recommender system scenario, we argue that recommender systems, also in the scenario of ads such as in (3), are bound to a limited domain or scope (i.e., limited to the underlying service's catalog, database, or index).

We see big potential in the use of novel applications that can be used to extend the knowledge and horizon of the user, present them with surprising and exciting new ideas, content, and information across different domains and modalities, and potentially burst filter bubbles that the user might find themselves in.

3.4.1 Risks and Opportunities of Pre-emptive IR

Pre-emptive IR seems like a promising direction for future research that does not come without risks. On the one hand, the opportunities of receiving information without the need for active information seeking might lower the threshold for users to explore a new field of interest and potentially even improve information literacy. On the other hand, deciding when to perform an intervention and present a user with information brings a delicate balance in understanding the user's cognitive availability, in addition to the high risk scenario of "predicting" relevance to a user, e.g., in the worst-case scenario we may present the user with unsollicited harmful content: which can be considered worse than presenting harmful content to a user that is already engaged in information seeking. Knowing the boundary of whether or not the user is open to receiving information of any kind might prove a difficult, yet interesting, task.

Future work could focus on questions such as what would we need to trust a system that gives us this information unprompted or how to provide this information (modalities, interface) and in which context (e.g., when you are on a laptop, or on the bus...?).

4 Final Note

The *Second 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. Despite the many open questions and challenges raised, the discussions throughout the workshop were marked by a shared sense of optimism. Participants identified new and evolving research directions, emphasizing a vibrant future for search and its role in a rapidly changing world.

Acknowledgments

We thank all the speakers and participants for their contributions, insights, and engagement throughout the workshop. We would like to thank Ian Soboroff (NIST) for their support in organizing the workshop. Their contribution is gratefully acknowledged. We are also grateful to the ECIR 2025 organizing committee for supporting our event and helping create a positive and memorable conference experience.

A Authors and Affiliations

Workshop organizers:

- Charles L. A. Clarke, University of Waterloo, Canada.
- Paul Kantor, University of Wisconsin Madison, USA.
- Adam Roegiest, Zuva, Canada.
- Johanne R. Trippas, RMIT University, Australia.
- Zhaochun Ren, Leiden University, The Netherlands.

Workshop participants:

- Maria Sofia Bucarelli (Sapienza University of Rome, Italy).
- Xiao Fu (University College London, UK).
- Yixing Fan (Institute of Computing Technology, China).
- Michael Granitzer (University of Passau, Germany).
- David Graus (University of Amsterdam, The Netherlands).
- Maria Heuss (University of Amsterdam, The Netherlands).
- Jaap Kamps (University of Amsterdam, The Netherlands).
- Yibin Lei (University of Amsterdam, The Netherlands).
- Andrew Parry (University of Glasgow, UK).
- Damiaan Reijnaers (University of Amsterdam, The Netherlands).
- Maarten de Rijke (University of Amsterdam, The Netherlands).
- Siddharth A.K. Singh (University of Amsterdam, The Netherlands).
- Yubao Tang (University of Amsterdam, The Netherlands).
- Suzan Verberne (Leiden University, The Netherlands).
- Jonas Wallat (L3S Research Center, Germany).
- Yumeng Wang (Leiden University, The Netherlands).
- Chen Xu (Renmin University of China, China).
- Andrew Yates (Johns Hopkins University, USA).
- Saber Zerhoudi (University of Passau, Germany).
- Jujia Zhao (Leiden University, The Netherlands).

References

Himan Abdollahpouri and Robin Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158*, 2019.

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare AI. *CoRR*, abs/2311.01463, 2023. doi: 10.48550/ARXIV.2311.01463. URL https://doi.org/10.48550/arXiv.2311.01463.

James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in Lorne. *SIGIR Forum*, 46(1):2–32, May 2012.

Avishek Anand, Abhijit Anand, Vinay Setty, et al. Query understanding in the age of large language models. *arXiv preprint arXiv:2306.16004*, 2023.

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. In *ACM SIGIR Forum*, volume 58, pages 1–41. ACM New York, NY, USA, 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.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. *Computational Linguistics*, 27(3):273–311, 2001.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety - A review. *CoRR*, abs/2404.14082, 2024. doi: 10.48550/ARXIV.2404.14082. URL https://doi.org/10.48550/arXiv.2404.14082.

Moumita Bhattacharya, Vito Ostuni, and Sudarshan Lamkhede. Joint modeling of search and recommendations via an unified contextual recommender (unicorn). In *RecSys*, pages 793–795. ACM, 2024.

Wenming Cao, Wenshuo Feng, Qiubin Lin, Guitao Cao, and Zhihai He. A review of hashing methods for multimodal retrieval. *IEEE Access*, 8:15377–15391, 2020.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *CIKM*, pages 191–200, 2022.

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, 2025.

Jeffrey Dalton and John Foley. Search agent model: A conceptual framework for search by algorithms and agent systems. In *Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, 2018.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *ICLR*, 2021.

Kalyanmoy Deb and Himanshu Gupta. Introducing robustness in multi-objective optimization. *Evolutionary computation*, 14(4):463–494, 2006.

Ashley Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019.

Juan Feng, Zuo-Jun Max Shen, and Roger Lezhou Zhan. Ranked items auctions and online advertisement. *Production and Operations Management*, 16(4):510–522, 2007.

Marc Formal, Camille Lassance, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292. ACM, 2021.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

Donna Harman. *Information Retrieval Evaluation*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2011. ISBN 978-3-031-01148-1. doi: 10.2200/S00368ED1V01Y201105ICR019. URL https://doi.org/10.2200/S00368ED1V01Y201105ICR019.

Mireille Hildebrandt. Qualification and quantification in machine learning. From explanation to explication. *Sociologica*, 16(3):37–49, 2022.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.386. URL https://doi.org/10.18653/v1/2020.acl-main.386.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1-2):1–224, 2009. doi: 10.1561/1500000012. URL https://doi.org/10.1561/1500000012.

Camille Lassance and Stéphane Clinchant. Mistral-splade: Efficient sparse retrieval with distilled representations. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1098–1110. ACL, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020a.

Patrick S. H. 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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. Interpretable multimodal retrieval for fashion products. In *MM*, pages 1571–1579, 2018.

Jimmy Lin, Pankaj Gupta, Will Horn, and Gilad Mishne. Musings about the future of search: A return to the past?, 2024.

Xinwei Long, Zhiyuan Ma, Ermo Hua, Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Retrieval-augmented visual question answering via built-in autoregressive search engines. In *AAAI*, 2025.

David Maxwell and Leif Azzopardi. Agents, simulated users and humans: An analysis of performance and behaviour. In Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi, editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 731–740. ACM, 2016. doi: 10.1145/2983323.2983805. URL https://doi.org/10.1145/2983323.2983805.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum*, 55:Article 13, 2021.

George A. Miller. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. Learning to relate to previous turns in conversational search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1722–1732, 2023.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. A survey of conversational search. *arXiv preprint arXiv:2410.15576*, 2024.

Jian-Yun Nie. Symbolic knowledge in neural information retrieval: Challenges and opportunities. *Journal of Information Retrieval*, 26(4):567–589, 2023. doi: 10.1007/s10791-023-09423-1.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework, 2024.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. *arXiv preprint arXiv:2406.13663*, 2024.

Jennifer E. Rowley and Frances C. Johnson. Understanding trust formation in digital information sources: The case of wikipedia. *J. Inf. Sci.*, 39(4):494–508, 2013. doi: 10.1177/0165551513477820. URL https://doi.org/10.1177/0165551513477820.

Yubao Tang, Ruqing Zhang, Zhaochun Ren, Jiafeng Guo, and Maarten de Rijke. Recent advances in generative information retrieval. In *The 46th European Conference on Information Retrieval*, 2024a.

Yubao Tang, Ruqing Zhang, Weiwei Sun, Jiafeng Guo, and Maarten de Rijke. Recent advances in generative information retrieval. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1238–1241, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400701726.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Shihao Liu, Shuaiqing Wang, Dawei Yin, and Xueqi Cheng. Generative retrieval for book search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Yi Tay, Vinh Q Tran, Mostafa Dehghani, Jianmo Ni, and Dara Bahri. Transformer memory as a differentiable search index. In *NeurIPS*, pages 21831–21843, 2022.

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.

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.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43 (2):55:1–55:37, 2025. doi: 10.1145/3708985. URL https://doi.org/10.1145/3708985.

Yumeng Wang, Xiuying Chen, and Suzan Verberne. Quids: Query intent generation via dual space modeling. *arXiv preprint arXiv:2410.12400*, 2024b.

Ofir Weller, Dawn Lawrie, and Benjamin Van Durme. Nevir: Negation in neural information retrieval. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2274–2287, St. Julian's, Malta, 2024. Association for Computational Linguistics.

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.

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, 2024a. Association for Computing Machinery. ISBN 9798400704314.

Chen Xu, Xiaopeng Ye, Jun Xu, Xiao Zhang, Weiran Shen, and Ji-Rong Wen. Ltp-mmf: Toward long-term provider max-min fairness under recommendation feedback loops. *ACM Trans. Inf. Syst.*, 43(1), November 2024b. ISSN 1046-8188.

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, 2025. URL https://arxiv.org/abs/2504.14991.

Lianwei Xu, Yichong Zhang, and Tao Li. Do large language models understand symbols? a critical examination of knowledge representation in pre-trained transformers. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, New York, USA, 2024c. PMLR.

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.

Hansi Zeng, Chen Luo, and Hamed Zamani. Planning ahead in generative retrieval: Guiding autoregressive generation through simultaneous decoding. In *SIGIR*, pages 469–480, 2024.

Saber Zerhoudi and Michael Granitzer. Searchlab: Exploring conversational and traditional search interfaces in information retrieval. In George Buchanan, Haiming Liu, Dana McKay, and Douglas W. Oard, editors, *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2025, MelbourneAustralia, March 24-28, 2025*, pages 382–389. ACM, 2025. doi: 10.1145/3698204.3716475. URL https://doi.org/10.1145/3698204.3716475.

Saber Zerhoudi, Michael Granitzer, Christin Seifert, and Joerg Schloetterer. Evaluating simulated user interaction and search behaviour. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II*, volume 13186 of *Lecture Notes in Computer Science*, pages 240–247. Springer, 2022a. doi: 10.1007/978-3-030-99739-7_28. URL https://doi.org/10.1007/978-3-030-99739-7_28.

Saber Zerhoudi, Michael Granitzer, Christin Seifert, and Jörg Schlötterer. Simulating user interaction and search behaviour in digital libraries. In Giorgio Maria Di Nunzio, Beatrice Portelli, Domenico Redavid, and Gianmaria Silvello, editors, *Proceedings of the 18th Italian Research Conference on Digital Libraries, Padua, Italy, February 24-25, 2022 (hybrid event)*, volume 3160 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022b. URL https://ceur-ws.org/Vol-3160/paper8.pdf.

Weinan Zhang, Junwei Liao, Ning Li, Kounianhua Du, and Jianghao Lin. Agentic information retrieval, 2025.

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.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024. URL https://arxiv.org/abs/2308.07107.