EVENT REPORT

Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025)

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

J. Shane Culpepper
The University of Queensland
Australia
s.culpepper@uq.edu.au

Mohammad Aliannejadi, James Allan, Enrique Amigó, Jaime Arguello, Leif Azzopardi, Peter Bailey, Jamie Callan, Rob Capra, Nick Craswell, Bruce Croft, Jeff Dalton, Gianluca Demartini, Laura Dietz, Zhicheng Dou, Carsten Eickhoff, Michael Ekstrand, Nicola Ferro, Norbert Fuhr, Dorota Glowacka, Faegheh Hasibi, Danula Hettiachchi, Rosie Jones, Jaap Kamps, Noriko Kando, Sarvnaz Karimi, Makoto P Kato, Bevan Koopman, Yiqun Liu, Chenglong Ma, Joel Mackenzie, Maria Maistro, Jiaxin Mao, Dana McKay, Bhaskar Mitra, Stefano Mizzaro, Alistair Moffat, Josiane Mothe, Iadh Ounis, Lida Rashidi, Yongli Ren, Mark Sanderson, Rodrygo Santos, Falk Scholer, Chirag Shah, Laurianne Sitbon, Ian Soboroff, Damiano Spina, Paul Thomas, Julián Urbano, Arjen de Vries, Ryen White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Justin Zobel, Shengyao Zhuang, Guido Zuccon[*]

Abstract

The purpose of the Strategic Workshop in Information Retrieval in Lorne (SWIRL)[1] is to explore the long-range issues of the information retrieval (IR) field, to recognise challenges that are on – or even over – the horizon, to build consensus on key challenges, and to disseminate the resulting information to the research community. The intent is that this description of open problems will help to inspire researchers and graduate students to address the questions and will provide funding agencies with data to focus and coordinate support for IR research.

Date: 10–12 February 2025.

Website: https://sites.google.com/view/swirl2025/home.

[*] Authors and participants (listed alphabetically). Affiliation not shown for all authors due to space limitations (see Appendix A for details).

[1] SWIRL 2025 was held in Torquay.

1 Introduction

Over the past twenty years, four Strategic Workshops on Information Retrieval (IR) have been organised in Lorne, Australia, all of which have had a singular vision – to look back at how research has evolved in the IR community and to look forward to where the research frontier is taking us. The first SWIRL workshop was organised by Alistair Moffat and Justin Zobel in 2004 and had 35 participants, including several PhD students. The major output of the meeting was the SIGIR Forum article "Recommended Reading for IR Research Students." [Moffat et al., 2005].

The second SWIRL workshop was organised by James Allan, Bruce Croft, Alistair Moffat, Mark Sanderson, and Justin Zobel in 2012. The theme of the workshop shifted away from previous work and focused more on future directions for the IR research community. Together, the 45 attendees debated several possible research topics and eventually converged on 6 main themes and 21 minor themes. These themes were then summarised and published in the SIGIR Forum article "Frontiers, Challenges, and Opportunities for Information Retrieval" [Allan et al., 2012].

In 2018, Shane Culpepper and Fernando Diaz organised the third SWIRL, bringing together 60 IR researchers from North/South America, Europe, and Oceania in Lorne to explore the future of IR. The central theme of the meeting was: *How has IR research evolved over the past five years, and where is it headed in the next five?* In preparation, participants completed surveys and homework assignments that helped shape discussions on key trends and challenges. Their report captures the insights gathered from the activities and summarises the key outcomes of SWIRL 2018, highlighting the evolving trajectory of IR research [Culpepper et al., 2018].

With the rise of generative systems, the playing field of IR has shifted, prompting the organisation of a new SWIRL workshop by Johanne Trippas and Shane Culpepper. A total of 60 researchers worldwide were invited to Melbourne to discuss the future of IR research and development. Just as in 2018, participants were asked to reflect on the evolution of IR by considering key questions: *What did previous SWIRL attendees accurately predict about the future of IR? What did they anticipate that did not come to pass? What major developments did they fail to foresee?* The fourth SWIRL aims to reassess past predictions, identify emerging challenges, and examine the evolving landscape of IR research, including the impact of generative AI (GenAI).

1.1 Workshop Format

SWIRL 2025 followed the format established in the 2012 and 2018 meetings. It began with an evening welcome reception where participants discussed insights from their homework assignments. The following day, a bus transported attendees from Melbourne to Torquay. After lunch, six seed talks were presented, setting the stage for deeper discussions. On the second day, participants split into six breakout groups of ten people to brainstorm the future of IR based on the seed talks. Each group identified and pitched key ideas, resulting in many overlapping proposed topics. These topics were then clustered by similarity, and attendees voted on the ones they were most interested in exploring further. In the afternoon, participants formed seven focus groups around the selected topics. These focus groups formed the core discussions of the workshop. Each topic has a section in this report (Section 2–Section 8), with additional "Minor Topics" included at the end (Section 9). The final day continued the focus group discussions, culminating in a collaborative effort to summarise key takeaways and produce a final report on the workshop's findings.

1.2 Invitation Questionnaire

As part of the initial RSVP for SWIRL, participants were asked what topics they thought were important. Table 1 shows the most common responses, with the number of respondents suggesting the topic is shown in parentheses. Unsurprisingly, "Generative IR and retrieval-augmented generation" is the most emphasised area. As per previous editions, substantial attention is also given to "evaluation and efficiency" addressing the need for robust evaluation metrics. "Human-AI collaboration and agents" and "IR scope and interdisciplinarity" are equally important, often linking to AI's role in automation and expanding the field's interdisciplinary reach. Other significant areas include user behaviour and personalisation, ethical considerations, and improving user interaction through human-centred design. The table also highlights the need for novel interfaces and content quality assurance alongside community-focused strategies regarding policy and promotion efforts.

Perhaps a more contemporary topic is the social impact of these advancements, particularly how GenAI reshapes information access, trust, and society. As AI-driven retrieval and generation become more prevalent, discussions around misinformation, bias, and societal dependencies on automated systems are gaining urgency. Additionally, responsible IR must consider diverse knowledge systems, including Indigenous knowledge, ensuring that AI respects, preserves, and ethically integrates these perspectives rather than marginalising them.

Table 1. Important topics suggested for the workshop agenda.

Topic	Count
Generative IR and retrieval-augmented generation (RAG) (Generative IR, RAG, GenAI, generative models for search and recommendation)	19
Evaluation and efficiency (evaluation of generative models, large language model (LLM) evaluation, evaluation of retrieval systems, bridging offline-online evaluation)	16
Human-AI collaboration and agents (task automation, human-AI cooperation, agentic systems)	15
IR scope and interdisciplinarity (broadening scope of IR, interdisciplinary approaches, IR & society)	14
User behavior and personalisation (user behaviour, personalisation, decision making)	11
Ethics, fairness, equity, and social responsibility (bias and fairness, responsible IR, green IR, Indigenous knowledge)	9
Human-centered information access (user experience, accessibility, enhancing user interaction with information systems)	8
Simulation, novel interfaces, and personal information management (simulations, novel interfaces, conversational search, multimodal IR)	7
Content quality and trustworthiness (content quality, creativity vs. hallucination, trustworthy outputs)	6
Community, policy, regulation, and promotion (policymaking and regulation, promoting the field)	5

1.3 Pre-Meeting Homework

1.3.1 Retrospective Questionnaire of Previous SWIRL Reports

The previous three SWIRLs produced reports, a list of recommended reading [Moffat et al., 2005], a vision document in 2012 [Allan et al., 2012], and a vision document in 2018 [Culpepper et al., 2018]. The homework task was to review previous SWIRL reports and answer the following questions:

1. What do you think previous SWIRL attendees accurately predicted about the future of IR (i.e., true positives: what did we get right)?
2. What do you think previous SWIRL attendees did not accurately predict about the future of IR (i.e., false positives: what did we get wrong)?
3. What do you think previous SWIRL attendees did not predict about the future of IR (i.e., false negatives: what did we miss)?

What did we get right? In Table 2, we present the recurring topics identified from previous SWIRL predictions, focusing on areas where the predictions were accurate. The table highlights the frequency with which each topic was anticipated. **Conversational information seeking** is the most prominent topic, appearing 21 times, reflecting a growing interest in the intersection of search systems and conversational interfaces. Closely following, **machine learning (ML) (neural) in IR** is noted 12 times, demonstrating the increasing role of neural-based techniques in IR tasks. **Evaluation**, appearing 11 times, highlights the ongoing importance of rigorous assessment methods for IR systems. The principles of **fairness, accountability, confidentiality, and transparency (FACT)** are mentioned 10 times, highlighting the rising concern with ethical considerations in IR research. The **generated information objects** category appears 9 times, emphasising the growing focus on creating and using automatically generated content. **Efficiency** is mentioned 7 times, continuing to be a key area of interest in improving system performance. Finally, **personalisation**, appearing 4 times, reflects the importance of tailored user experiences in search systems.

Table 2. Recurring topics from previous SWIRL predictions (i.e., what did we get right?).

Topic	**Count**
Conversational information seeking	21
ML (Neural) in IR	12
Evaluation	11
Fairness, accountability, confidentiality, and transparency (FACT)	10
Generated information objects	9
Efficiency	7
Personalisation	4

What did we get wrong? The reflections on the second question (i.e., *what did we get wrong?*) provide a nuanced critique of the past SWIRL predictions, highlighting successes and shortcomings in anticipating the future of IR. While many attendees indicated that the predictions captured emerging trends, the predictions often overestimated the progress or failed to foresee the rise of disruptive technologies that would shift the field of IR.

We summarise and highlight some reflections:

- **Technological advances and shifts:** One significant change is the rise of large language models (LLMs) and their impact on IR, particularly in conversational search, relevance assessments, and fairness. Previous reports did not fully anticipate these models, but their influence is definite. The failure to predict LLMs' centrality in current IR applications highlights a blind spot in earlier forecasts. Agentic AI and its potential to handle complex search tasks have emerged as a new frontier, though they were not fully predicted.
- **Underestimated challenges and overestimated impacts:** Many predictions concerning voice-based search, personalised search, and decision support systems proved overly optimistic. Challenges related to privacy, data availability, and the technical complexity of these systems have slowed their adoption and impact in academia and industry. Predictions about the widespread adoption of new evaluation paradigms and personalisation techniques were also not fully realised. Similarly, while multi-stage search systems and efficiency were expected to be major topics, their progress has been overshadowed by deep learning and domain-specific retrieval, where model effectiveness often takes precedence over efficiency.
- **Unfulfilled evaluation:** Despite significant advancements in evaluation, particularly in diversity, fairness, and counterfactual analysis, a clear link between offline and online evaluation remains elusive. The expected breakthroughs in evaluation paradigms have not materialised, highlighting the continuing need for improved metrics to assess the real-world effectiveness of IR systems.
- **Underestimated societal impact of information access:** Earlier SWIRL reports gave limited attention to societal impacts, such as fake news and election interference, which have become more pressing. Although FACT concerns were part of past discussions, they have gained significant attention in the AI community and are now increasingly relevant to IR. Additionally, the rise of AI-generated "hallucinations" in systems like LLMs has introduced challenges in trust and reliability, further complicating the relationship between AI and IR.

There is an overall sense that while the past SWIRL reports did not predict every shift in the field, they were still mainly in the right direction. The future of IR is shaped by LLMs, Agentic AI, and increasingly sophisticated systems that aim to address more complex tasks. However, many foundational issues, such as personalised search and fairness, still need further exploration and resolution.

This reflection by the SWIRL participants highlights IR's dynamic nature. It shows how emerging technologies like LLMs have disrupted earlier predictions. It also points out areas where predictions missed the mark or failed to anticipate challenges that slowed progress. It emphasises the importance of focusing on technological progress and considering the broader social effects and challenges that come with these advancements. Perhaps future predictions about IR should be more adaptable and consider both the tech side and the impact on society.

What did we miss? There is a consensus that previous SWIRL reports greatly underestimated the rise of LLMs as seen in Table 3. Emerging technologies like LLMs, GenAI, and RAG were notably absent from earlier predictions despite their influence on the field in recent years. Additionally, trends such as the convergence of IR, natural language processing (NLP), recommender systems, and other ML-driven fields, the increasing role of LLMs in evaluation, and their societal impact were overlooked.

Table 3. Topics not predicted by previous SWIRL reports (i.e., *what did we miss?*).

Topic	**Count**
LLMs	19
GenAI and RAG	16
Convergence of IR and closely related fields	8
Multimodality and crossmodality in IR	7
Sustainability and green IR	6
Role of LLMs in evaluation	5
Agentic IR	4
Impact of LLMs on societal aspects	4
Scalability and efficiency challenges	3
Recommender systems integration	2

Other less frequent topics include: *(i)* quantum and quantum IR, *(ii)* political interferences, *(iii)* political interferences, *(iv)* economic aspects, *(v)* responsibility of IR, *(vi)* user simulation, *(vii)* synthetic data generation, and *(viii)* reasoning.

1.3.2 Important Paper Topics Since SWIRL 2018

As part of the homework assignment, we asked participants to select one paper from their area of expertise and another from outside their field that they felt was important for the IR community. We then manually classified these papers to understand their recent research perspectives better, the overview of the topics are presented in Table 4.

Table 4. Important paper topics since SWIRL 2018 (from participants' core expertise).

Topic	**Count**
LLMs in IR	10
Evaluation	4
Users and information	3
Dense retrieval	3
RAG	2
Bias	2

1.4 Summary of Seed Talks

Based on the RSVP questionnaires and homework assignments, "fire starter" talks were proposed to engage participants' interests and stimulate provocative discussions. In addition, overviews of recent workshops focusing on future-looking directions in IR were presented to provide context on trends and challenges. The topics were:

LLMs. IR is in a period of rapid change – whether that leads to the extinction of IR as we know it, or not, is an open question - is IR dying? LLMs and GenAI are driving much of this change – enabling and creating new ways in which people (and machines) engage and interact with information. Human traditional search engine usage has peaked, with AI agents becoming the main users. This is because humans are increasingly expecting to have more seamless, fluid, and dynamic interactions (and conversations) with and over the information space rather than traversing documents. This shift marks a move to an agentic world, where information is distributed across networks and accessed by other agents. In this multi-agent setting (think federated IR on steroids), users will expect even more of information (retrieval) systems. They will want to not only fulfil their information needs but also to explore, learn, (co)create, and complete tasks through controllable, explainable and bespoke conversational interfaces tailored to their contexts.

RAG. RAG is an emerging research direction at the intersection of LLMs and IR, offering key benefits like access to current information without retraining, evidence-based generation, and efficient context handling. RAG enables secure applications such as enterprise search by separating general-purpose models from private or domain-specific data. However, major challenges remain, including unclear interactions between internal knowledge and retrieved content, addressing bias and fairness in retrieval and generation, and adapting traditional IR methods (i.e., indexing and evaluation). Economic trade-offs and the integration of user feedback across multiple turns also remain open research problems. RAG ultimately calls for a closer integration of IR and NLP to build more reliable, adaptive systems.

Automated relevance labelling. Automated relevance labelling using LLMs has emerged as a prominent topic, provoking debate on whether LLMs should replace human assessors, complement them, or be avoided entirely due to risks such as circularity, biases, and data contamination. A balanced approach suggests shifting from competition towards collaboration, clearly defining roles for humans and models in hybrid agent-based assessments. Looking ahead five years, opinions diverge: some argue that automated labelling will never replace humans because of inherent risks and unknown biases, while others suggest that superhuman performance is achievable by self-improving systems (e.g., AlphaGo Zero). Automation could support scenarios like personalised relevance judgements ("Infinite Qrels") tailored to personas or privacy-preserving evaluations when humans cannot review sensitive documents. Despite these opportunities, meta-evaluation remains essential for identifying biases and ensuring evaluation reliability beyond traditional Cranfield methods. Ultimately, we must maintain scepticism yet openness regarding automated labelling's potential; human input likely remains the gold standard for IR evaluation, but may be complemented by robust automatic procedures.

Multimodal IR. Multimodal and cross-modal retrieval have existed for decades. Recent trends, especially the rise of short-form videos on platforms like TikTok, highlight the need for IR systems that can effectively process, integrate, and retrieve information across diverse modalities. Traditional text-based retrieval methods fall short as user behaviour shifts toward visually and aurally rich platforms. This evolution demands new models capable of multimodal reasoning, efficient indexing of complex content types (especially video), and retrieval strategies that account for explicit queries and implicit signals such as tone, emotion, and context. Deeper collaboration between the IR, vision, and language communities is essential to address these issues. Developing effective multimodal foundation models requires shared efforts in model design, evaluation, and user interface research. The goal is to improve retrieval accuracy and support meaningful, trustworthy, and engaging interactions. As content consumption habits evolve, IR research must adapt, building systems that reflect how users seek and process information in today's multimodal digital environments.

Impact on society and environment. Information access systems profoundly impact society, shaping what people see, how they make decisions, and who has access to opportunities. However, these systems also introduce risks: filter bubbles, biases, manipulation, and harm. In addition, their environmental costs (e.g., energy and water consumption) raise further concerns about who bears the burden and who benefits. To address these issues, we must ask: who is harmed by IR systems, how are they harmed (through exclusion, misrepresentation, or denial of access), and who controls the design and deployment of these technologies? A more honest approach requires shifting power in how IR systems are built and used. Emancipatory IR offers such a framework by opposing surveillance and manipulation and designing systems that support equity, sustainability, and collective well-being. This means moving beyond isolated "IR for good" projects toward a clear vision of the societal outcomes we want and a plan for how IR can help achieve them. Progress starts with collaboration. We need interdisciplinary spaces that bring together IR researchers, social scientists, and justice-focused practitioners to examine harms and critically guide the design of fairer systems.

IR System Users: New Directions. The presentation argued that IR research needs to refocus on real users rather than relying primarily on algorithms or simulated behaviours. There has been a shift of user-centred research to forums like CHIIR, but user studies should remain central to IR due to the importance of understanding persistent and evolving user behaviours, particularly in systems incorporating LLMs and GenAI. Users need to be viewed as complex individuals immersed in ongoing information environments, not as isolated actors engaging with IR systems in artificial ways. This could be facilitated by moving beyond query-centric models, with research focusing on users' tasks, experiences, and contexts. The talk called for IR research to support seamless and effective interaction with information, moving beyond traditional system boundaries.

Query Performance Prediction. Query performance prediction (QPP), estimating retrieval effectiveness without relevance labels, remains an essential research area, even as LLMs are reshaping the IR landscape. Recent developments show promise, including embedding-based models such as SPLADE and LLM-based predictors on benchmarks like MS MARCO. However, whether these approaches address the QPP's core challenges or only offer marginal gains is unclear. This uncertainty points to a deeper issue: the need to define QPP's problem space and objectives in modern retrieval paradigms. Traditional evaluation methods, often based on correlation metrics like Pearson or Kendall, are known to be sensitive to ties and outliers, limiting their robustness and interpretability. As a result, more reliable and reproducible evaluation strategies are needed to reflect practical performance differences and theoretical validity better. Moreover, QPP research must increasingly account for real-world system constraints. For instance, latency remains a critical factor, with evidence from industry (e.g., Google) showing that delays as small as 400 milliseconds can significantly reduce user engagement, highlighting the importance of incorporating efficiency into future QPP benchmarks alongside predictive accuracy. We must return to foundational principles through systematic experimentation and theoretically grounded, axiomatic approaches. This involves rethinking how we evaluate QPP methods, why we pursue them, and their role in the broader information access ecosystem. The field now faces a pivotal choice: continue refining existing tools incrementally or fundamentally reimagine QPP.

The Many Dimensions of Efficiency. This talk was motivated by the IR community's strong focus on system building and the growing recognition that neglecting efficiency can lead to unsustainable computational costs. This concern is particularly acute in modern LLM-powered indexing, ranking, and retrieval systems, which are already resource-intensive and are likely to become even more so as data volumes grow and models scale further. The central argument is that efficiency must be treated as a first-class design objective, not an afterthought. Importantly, efficiency encompasses more than just speed; it includes latency, throughput, memory usage, energy consumption, and environmental impact (e.g., carbon emissions). Efficiency also takes on new meaning in emerging contexts as machine-generated content proliferates. We must rethink how we retrieve and filter relevant information and anticipate how shifts in hardware and interaction paradigms will redefine what efficient access looks like. Addressing these challenges requires a holistic view of efficiency that aligns with technical scalability and responsible computing.

Future of IR Research in the Age of Generative Artificial Intelligence [Allan et al., 2024]. Recent discussions on the short- and long-term research directions combining IR with GenAI led to the identification of key research areas. These include developing interactive IR and GenAI systems that cooperatively determine what information to retrieve and leveraging explicit and implicit user feedback. Emphasis was also placed on constructing task-agnostic foundation models capable of multimodal, personalised, and continuous learning. Creating efficient, personalised AI "digital twins" for recommendation and information synthesis tasks was a significant goal, particularly when paired with transparent, persuasive explanation mechanisms. Mixed-initiative agent systems were also highlighted as promising, requiring new evaluation frameworks to accommodate multi-agent collaboration, proactive behaviour, and challenges such as hallucination. Computational trade-offs were discussed, including comparisons between smaller and large LLMs, the design of fixed-footprint models, and the implications of emerging computing paradigms such as quantum computing. Evaluation methodologies were recognised as requiring revision, moving beyond binary relevance to multi-step conversational search, response accuracy, simulated user studies, and explicit explanations. Future work should also focus on modelling multimodal user interactions using privacy-aware techniques. Ultimately, ensuring the responsible use of GenAI will require ongoing collaboration with social scientists, legal experts, and policymakers to address societal impacts.

A view from the Chinese IR community [Ai et al., 2023]. Recent LLM advances are pushing the IR community to revisit core retrieval concepts. The definition of a corpus is expanding beyond documents and webpages to include microblogs, dialogues, and multimodal knowledge sources. Retrieval is evolving from basic search to more complex tasks such as recommendation, summarisation, and question answering. Users are no longer seen as passive recipients but active contributors within interactive systems. Promising research directions include neural embeddings for semantic representation, reinforcement learning for long-term user modelling, and decentralised architectures for distributed data environments. These shifts introduce generalisation, interpretability, and evaluation challenges, calling for new frameworks that reflect spatial and temporal context, user personalisation, and transparency. Strategic efforts, including those by the Chinese IR community, stress IR's unique strengths in intent modelling, ranking, and application in medicine, education, and legal advice. As LLMs increasingly rely on retrieval for grounding and factual consistency, IR is positioned to play a central role in enabling reliable, efficient, and context-aware generative systems.

1.5 Summary of Brainstorming Breakout Sessions

Six breakout groups built on the seed talks by exploring existing themes and introducing new ones they felt were missing. Some topics reinforced earlier discussions, while others expanded the conversation in fresh directions. Their combined insights shaped the following key themes:

- **RAG & generative models:** RAG, generative IR models, multi-document summarisation using GenAI
- **Efficiency & scalability in LLMs:** Effectiveness-efficiency trade-offs, computational efficiency concerns related to LLM deployments
- **Evaluation methodologies for GenAI in IR:** New metrics/frameworks specifically designed for generative systems, multi-aspect/multimodal evaluations for RAG
- **Traditional evaluation approaches & reproducibility standards in IR:** Offline vs online evaluations, reproducibility standards/reporting guidelines, bridging offline-online evaluation divides
- **User-centric interactive search systems:** Conversational search user experience, interactive IR interfaces, novel user interfaces/interaction design
- **Personalisation & user modelling techniques:** Personalised search/recommendation hybrid systems, user modelling simulations, personal information management (PIM)
- **Human-AI collaboration & hybrid decision making:** Human-AI cooperation frameworks/models, collaborative search behaviours/approaches, retrieval-augmented decision-making/planning
- **Fairness, bias mitigation & ethical concerns in IR/AI:** Algorithmic fairness/bias detection & mitigation strategies, ethical dataset practices/authority veracity concerns
- **Explainability, transparency & trustworthiness methods:** Neuro-symbolic/explainable techniques, transparency frameworks/hallucination detection approaches
- **Diversity, equity, inclusion & decolonisation efforts in IR research community:** Indigenous voices/perspectives inclusion, diversity, equity, and inclusion initiatives within research community practices, decolonising IR research methodologies
- **Multilingual and low-resource languages IR:** Cross-lingual/multilingual IR approaches, low-resource language support techniques
- **Multimodal & contextually adaptive information access systems:** Multimodal retrieval (vision-language-audio modalities integration), context adaptation/context-aware retrieval systems
- **Sustainability IR:** Environmental sustainability ("green computing") issues, energy-efficient training/deployment of large models

- **Expanding scope:** Broader definition beyond document retrieval, integration with NLP/recommendation systems/decision-support tasks and interdicplinary work, debates about renaming/rebranding the "IR" discipline

1.6 Summary of Focus Group Breakouts

A poll was held for participants to identify the most interesting topics. This resulted in seven topics, which formed the final breakout focus groups. The focus groups discussed their topic and developed the summary reports found in the following sections:

- **Section 2:** Efficiency, scalability, cost, and sustainability
- **Section 3:** GenIA: Foundations of Generative Information Access Models and Systems
- **Section 4:** What is IR in New User Experience, Information Access & Interaction Scenarios?
- **Section 5:** Agentic Information Retrieval (AIR): IR for All
- **Section 6:** LLM-based Simulation for Evaluation
- **Section 7:** Centering Societal, Democratic, and Emancipatory Values and Ethics in IR
- **Section 8:** Evaluation of Complex IR

In addition to the main themes, participants highlighted eight topics they felt were important to discuss in greater depth. These are presented separately in Section 9. While they do not follow a shared structure like the main themes, each is addressed individually through key research questions and challenges. This approach reflects the workshop's recognition of their distinct importance and the need for focused discussion.

2 Spend Less, Get More: Efficiency, Scalability, Cost, and Sustainability

2.1 Description

The IR community has long been concerned with system efficiency due to the vast scale of data and users. Efficiency encompasses various aspects such as system efficiency (compute, storage, memory, network costs), data efficiency (labelling/annotation), engineering efficiency (development, maintenance), and user efficiency (effort, frustration). The advent of complex technologies that enhance information access experiences prompts questions about end-to-end task support and holistic cost measurement. Like human-computer information retrieval and other closely related fields, IR emphasises user-focused efficiency gains and reducing resource consumption. The IR community must consider multiple efficiency factors, even if individual practitioners focus on specific aspects only.

Current research in IR now has the opportunity to revisit and expand our understanding of efficiency concepts by developing new methodologies and creating tools to measure "cost". Crafting new measurement frameworks is challenging due to the many dimensions of efficiency and the potential tradeoffs. The field could measure efficiency as the total operational cost or map costs to metrics such as electricity consumption or CO_2 emissions. While kWh and CO_2 emissions may be common units to measure system efficiency, variability and other non-energy requirements must be considered alongside traditional metrics such as storage and latency. Discussions at SWIRL highlight the need for better frameworks to guide community thinking about efficiency as well as the standard practices used to measure system costs and efficiency gains in the IR field.

2.2 Motivation

Efficiency has always been a core interest within the IR community and continues to be a focus area today. Operation at the scale required by production search systems is only possible if efficient implementations are available, and that nexus will only get stronger as data volumes continue to grow. The transition to complex AI-based systems has translated into increased costs, making it increasingly important to control the constraints associated with the required data volumes and access operations. Developing efficient systems enables the democratisation of technology and prevents a monopoly by a handful of entities that can afford to develop and run resource-intensive IR models.

2.3 Proposed Research

We need fundamental improvements to core algorithms, representations, and architectures, considering exact or approximate solutions, hardware choices, and alternative meanings of efficiency. The key topics to consider are:

- **Measurement:** Optimising metrics to measure efficiency is crucial, requiring accurate information capture and standard evaluation methods supported by infrastructure such as code libraries for energy consumption reporting.
- **Standardised environments:** Controlled hardware environments are necessary for efficiency benchmarking, with consistent measurement practices, modelling a range of operating environments.
- **Adaptive and resource sensitive approaches:** Efficient architectures should adapt to tasks within an efficiency budget, using resource-economical methods and improved benchmarks with context.
- **Efficiency at the edge:** Local processing enhances privacy and performance, allowing local and global data integration, with potential for new interfaces and real-time efficiency considerations.
- **New-generation hardware:** Consideration of new and bespoke hardware, like quantum computing and Language processing units, is essential for improving IR, with IR experts playing a crucial role.

2.4 Research Challenges

Efficiency has become a more complex topic since SWIRL 2018 due to the substantial increase in the computational complexity and data requirements of recent information access systems concurrent with increased attention paid to the carbon costs of technology. The efficiency breakout group organised its discussion of efficiency-related research challenges into the four broad topics described above: Computational efficiency, data efficiency, engineering efficiency, and user efficiency. Each of these factors contributes to the total cost of developing and providing any particular retrieval or question-answering service and needs to be carefully balanced against the net social benefit that accrues from that service. The discussion related to each factor is summarised below.

2.4.1 Computational Efficiency

The last few decades have witnessed significant advancements in improving efficiency in IR systems, focusing primarily on metrics such as time to create inverted indexes, storage space, and query resolution speed. These improvements are crucial for handling large volumes of data and meeting user expectations for quick responses. The SWIRL 2025 discussion primarily considered the ongoing importance of these core technologies and the need for new methodologies to tackle emerging challenges.

Environmental Concerns. The IR community must address environmental responsibilities by considering the use of scarce resources and greenhouse gas emissions. Researchers should be encouraged to report electricity costs alongside traditional metrics like CPU hours and elapsed milliseconds, as these provide a better understanding of the real environmental impact, including water consumption and hardware fabrication costs.

Total Cost of Operation (TCO). A deployed system's total cost involves more than just the individual components used to build it. It includes data acquisition, storage, index construction, and query resolution. Energy costs should be amortised across these activities, and consider dependencies between components. For instance, faster index construction might lead to slower query resolution but could still be cost-effective if system updates happen frequently.

Specialised Hardware and Caching. The increasing importance of GPUs and the potential for more collaboration between hardware designers could also reduce IR system costs. As IR systems become increasingly more complex and personalised, traditional caching methods must be rethought. However, opportunities like the KV cache in current transformer-based architectures could promote better data reuse. As system architectures evolve, new challenges and opportunities will continue to arise, with system components competing even more for the limited cache resources available in current hardware.

2.4.2 Data Efficiency

LLMs are developed using vast amounts of data, which makes them expensive to develop and difficult to create for organisations that are less expert or do not have access to massive volumes of documents, synthetic training data, or user interactions (see also User Efficiency, below). This approach may also be unsustainable if model developers are eventually required to get permission from or compensate copyright owners. Synthetic data generation has emerged as an important tool for training models, but it is not free and requires a clear understanding of the properties needed in the synthetic data. A significant opportunity exists if equally powerful models can be trained and updated from smaller, better-curated corpora with less synthetic data and fewer user interactions. Progressing on this topic requires a deeper understanding of how large models learn, how they store information, and the sensitivity of different model properties to the amount, type, and quality of training data. In particular, it requires a deeper understanding of emergent properties observed so far only in very large models.

2.4.3 Engineering Efficiency

The academic community often overlooks the cost of engineering and maintaining complex IR systems, which includes financial costs, effort, skill, and expertise in research and development. Simpler pipelines with fewer components generally require less maintenance effort than complex ones. One recommendation is to consider the return on investment of new features alongside their expected lifespans. Although the industry is aware of these issues, the academic community has paid limited attention. The academic IR community should learn from industry practices and develop better awareness, creating a shared vocabulary for discussion and acknowledging the costs involved in developing and maintaining open-source research tools and artefacts.

2.4.4 User Efficiency

User efficiency can be considered from two main perspectives: informing users about efficiency costs and measuring the efficiency of user tasks. Informing users involves communicating about efficiency metrics, such as sustainability metrics, which can influence user decisions. For instance, search engines and information access services like Google, Bing, ChatGPT, and DeepSeek could report the "cost per query" in terms of watts or star ratings, similar to energy ratings for appliances. Such transparency could drive the construction of more efficient systems and enable more innovative models that require the same resources or make these systems accessible on lower-cost devices, promoting equality and democratisation. Kate Crawford's suggestion to consider factors such as litres of water per GPT transaction would highlight the broader environmental impact of these technologies.

Measuring user task efficiency continues to be less explored than measuring relevance. Industry practitioners with direct access to large user populations have led in characterising and comparing end-to-end task efficiency. This involves examining efficiency and effectiveness, which have complex interactions that depend on the system design. The human-computer interaction (HCI) and human-computer information retrieval communities have expanded the understanding of efficiency beyond measuring only runtime by considering the cognitive load that affects outcomes such as user satisfaction or frustration. However, conducting user studies over variations in system design is challenging, where the cost can be prohibitive when one must consider the impact of each component in increasingly complex systems. Using synthetic user personas created using generative models could help automate and expand these investigations, providing deeper insights into user task efficiency at a much lower cost.

2.5 Broader Impact: Influence on Other Research Fields and Society

The IR community has historically focused on balancing effectiveness and efficiency when building user-centric systems. This focus has led to the development of experimental methodologies, algorithms, and measurement approaches considering system performance and user experience. The need to manage growing data collections with limited hardware has driven innovation in efficiency and scalability. Other computing disciplines, such as NLP, data mining, ML, and distributed computing, have developed their strategies to address these challenges. The IR community can benefit from more interdisciplinary collaborations, workshops, and publications that share insights on building user-centric systems that emphasise the importance of empirical evaluation that balances effectiveness and efficiency.

Collaborating with colleagues in the human-computer information retrieval sub-discipline can expand our understanding of user effectiveness and efficiency. As interactive information systems evolve, they will increasingly integrate technologies from NLP, AI/ML, and related fields with IR systems. Collaborative efforts across these complex systems are essential for effectively communicating efficiency considerations to providers and customers. Additionally, IR systems can inform users about the environmental costs of their search behaviour, offering opportunities for microeconomists to study behavioural changes and inform policymakers.

2.6 Broadening the IR Community

Aside from engaging with the broader system efficiency research communities (e.g., energy informatics, hardware and operating systems), measuring the efficiency of an IR system creates opportunities for engaging with the HCI community. Users need to be involved in the estimation of efficiency measures in two ways: Computational efficiency: each user-system interaction incurs computational costs; a computationally efficient system that requires multiple user-system interactions to satisfy a user's information need may consume more resources than a more computationally expensive but also more effective system; and System efficiency: measuring user task efficiency with respect to time, effort, and other utility factors such as cognitive load. The HCI field can give the IR community insights into user behaviour and interaction patterns. These insights can be further used to develop user interaction simulators to optimise IR systems for both efficiency and effectiveness measures.

2.7 Obstacles and Risks

There are several challenges in establishing a standard set of metrics and methodologies to measure efficiency, which can complicate experimental comparisons of new approaches to search (and recommendation). Conducting fair comparisons requires significant effort and access to a common reference infrastructure, which can be hindered by costs and/or institutional policies. Additionally, achieving genuine scale in experimental comparisons in research remains a challenge due to the increasing cost and complexity of IR pipelines despite the availability of large-scale open-source datasets. There is also a risk that efficiency improvements are overlooked in other closely related fields, such as electrical engineering and quantum computing, emphasising the need for a broader perspective on efficient algorithm design in the IR domain.

3 GenIA: Foundations of Generative Information Access Models and Systems

3.1 Description

The IR community has traditionally focused on discrete information representations, such as individual words and documents, with keyword queries as input and ranked lists of documents as output. However, the rapid evolution of GenAI techniques has prompted us to rethink information access technologies. Users now expect more holistic search results, including direct answers to questions, multimodal content, complete and correct coding solutions, and even transactions performed on their behalf. This shift challenges traditional IR systems, which rely on specialist components such as crawlers, tokenisers, and statistical retrieval models. While foundation models may eventually replace some of these heuristic approaches, they may introduce new concerns and blur the lines between various information access tasks, such as recommendation systems and question-answering.

The emergence of generative information access systems raises questions about the essential building blocks of IR systems, as there is no common language to describe their constituents and composition. Generative information access systems extend beyond keyword requests and document lists, lacking basic terminology for new input/output objects. This section in the SWIRL report will explore modern information access solutions' design objectives and concrete implementations, identifying frameworks, components, and challenges for future systems. The landscape of information access is being transformed by technological change, necessitating a reevaluation of established truths and the development of frameworks and components suitable for the age of GenAI.

3.2 Motivation

We are entering a new phase of intelligent information systems driven by recent advances in AI and generative foundation models. This new era will transcend traditional device constraints and interaction modalities like speech, images, and video to create innovative and intelligent information access models. These systems will support various capabilities, including generative, conversational, multimodal, and stateful interactions, all designed with rich human and agent interaction protocols. Unlike the current models, which integrate retrieval and AI systems at the component level, **the future architecture will be built from the ground up with foundational models specifically designed for information access**. This involves rethinking every step of foundation model design, from new forms of pretraining that incorporate search to post-model training and fine-tuning with generative retrieval.

The new system architecture will exhibit several key properties. It will be generative, capable of producing rich outputs, and interactive, engaging in conversational exchanges across various modalities. The system will adopt a task-oriented approach, supporting multi-turn interactions for complex tasks. It will be multimodal, accommodating inputs and outputs through voice, images, gestures, and potentially brain interfaces. The system will be adaptive, contextual, and personal, adjusting to the environment and user emotions. Transparency will be a priority, with the system explaining its behaviour and reasoning about biases. It will be scalable and dynamic, capable of evolving with new information and feedback. The architecture will also be communicative, immersive, compliant with ethical and legal standards, and reusable, ensuring flexibility and adherence to governance and regulatory standards. These properties will fundamentally reshape the system architecture, whether a large or modular hybrid model.

3.3 Proposed Research

***(i)* System architecture design.** Recent advancements in LLMs are driving new retrieval model designs, raising important questions about which foundation models will be optimal for information access architectures. Key considerations include pre-training, instruction-tuning, and alignment data for system architectures. The concept of a "document" or "information item" needs to be reconsidered by exploring various granularities and latent information items for effective access. Both end-to-end and modular search engine architectures should be studied, focusing on efficiency, effectiveness, and robustness. Future architectures should enable adaptable, dynamic, and instructable access to information, with transparency being crucial.

***(ii)* Multimodal foundation models.** Future information access systems must leverage advancements in foundation models, particularly in large vision, language, and multimodal models. A shared generative model should support diverse tasks like recommendation and question-answering while balancing memorisation and generalisation. Developing unified frameworks for multimodal representation learning is essential, harmonising heterogeneous data types and integrating non-traditional modalities like brain-computer interaction signals. Effective information access demands optimising the content structure and enabling real-time updates without losing important information ("forgetting").

***(iii)* Personalisation, memorisation, and contextualisation.** The future of IR lies in systems that dynamically adapt to individual users through personalisation, memory, and contextualisation. Challenges include learning user interests, integrating short-term and long-term interactions, and addressing cold-start scenarios. Memory and context representation are crucial, ensuring user-specific information remains updatable and relevant. Architectural design decisions will shape personalised IR, balancing private and public information while addressing privacy and forgetting challenges (see above).

***(iv)* Sensing, reasoning, and actuation.** Modern IR systems require advanced intent understanding processes to tailor responses to user contexts. Dynamic architectures should selectively use suitable components, deploying collaborative agents for tasks like query decomposition and rewriting. Systems should be self-reflective, continuously learning from interactions to enhance performance. Multi-document reasoning and proactive approaches are essential, moving from finding information to identifying gaps and assisting users in accomplishing tasks.

***(v)* Reinforcement learning from user interactions.** User interaction data is crucial for evaluating and training IR systems. Advanced generative information access systems introduce dynamic user interactions involving direct feedback (e.g., explicit ratings or responses) and more subtle forms of engagement (e.g., scrolling behaviour, mouse movements, or navigation patterns). Research should focus on mixed-initiative interactions, optimal presentation modes, and evaluating multi-turn conversations. By analysing these direct responses and implicit feedback, generative information access systems can be further improved. In addition, digital twins and proxy agents offer mechanisms for collecting such data and providing real-time adaptive feedback to enhance user experience.

(vi) **Cost-aware, efficient, and adaptive architecture in low-resource contexts.** Low-resource environments require lightweight architectures, balancing accuracy and efficiency. Innovations in system architecture and strategies are needed to address data scarcity and heterogeneity. Cross-modal and multilingual data augmentation should be prioritised. Systems must dynamically adjust computational complexity based on query priority, deploying edge-cloud hybrid systems for cost efficiency and real-time responsiveness.

(vii) **Transparency, trustworthiness, and compliance.** GenAI in information access systems face many challenges, such as hallucination and citation accuracy, affecting transparency and trustworthiness. Research in attribution and grounding is already underway, with verifiable AI methods needed for dependable engineering. Accountability and compliance are crucial, with regulations proposed worldwide. System creators must provide compliance guarantees, addressing platform governance and decision-making interests.

3.4 Research Challenges

The research agenda for information access architectures is ambitious and interconnected to many related topics, necessitating a community effort to address key challenges. For example, a major challenge is securing high-performance computing resources, necessary for training advanced models but often unavailable to smaller research groups due to financial or institutional limitations. Shared infrastructure and collaborative partnerships can help address this gap. Large-scale model training and inference also negatively impact the environment. Researchers must prioritise sustainable computing strategies, such as improving energy efficiency, using renewable energy sources, and developing efficient model compression techniques. These approaches will help ensure that cutting-edge information access technologies are accessible and environmentally responsible.

Evaluation methodologies and infrastructure are insufficient, with a lack of large-scale publicly available data. Comprehensive corpora covering diverse cultures, languages, and media types are essential for training models and demonstrating research impact. Sharing such data involves practical and legal complexities, especially concerning user data, which requires restrictive conditions and extensive ethical reviews. Ensuring generality in models is challenging due to biases towards dominant languages and user profiles. Research must focus on generalising results to all information access scenarios and involving diverse stakeholders in architectural design decisions.

Compliance with ethics and regulatory frameworks, such as the European General Data Protection Regulation (GDPR) and the AI Act, presents technical challenges for information access architecture. Ensuring architectures comply with these frameworks involves handling personal data in training and user interactions, addressing memorisation of personal data, and implementing safeguards like differential privacy. Techniques such as model editing may be necessary to implement rights like the *right to be forgotten* or the *right to erase.* Full training provenance and effective unlearning of information are crucial to avoid costly retraining.

Addressing these challenges requires a concerted effort from the research community to develop sustainable, inclusive, and compliant information access architectures. By leveraging collaborative resources, refining evaluation methodologies, and adhering to regulatory frameworks, the community can advance the field while ensuring ethical and equitable access to information.

3.5 Broader Impact: Influence on Other Research Fields and Society

Generative information access place IR at the forefront of the current AI revolution, driven by LLMs and AI. Historically, IR has been central to AI, with its roots predating the 1956 Dartmouth Conference. The field's evolution has converged with NLP, where IR's influence is evident in adopting vector-space models. Rather than one field overtaking the other, current NLP methods build on classic IR principles. IR also intersects with social sciences, information science, HCI, ethics, and behavioural research. Both system-focused engineering and user-centred design are critical for understanding interactive search behaviours. As generative technologies change how users interact with information systems, researchers must examine technical performance, user experience, and societal effects. The socio-technical impact of generative information access tools is profound, changing user behaviour and practices. The field is experiencing a major shift similar to the one brought about by the Cranfield experiments in the 1960s. Generative models are changing how users interact with search and how they synthesize knowledge and make decisions. These systems will evolve as users adapt their behaviours to new capabilities. Researchers must make careful architectural choices so that these systems remain adaptable rather than settling into inefficient patterns.

Business priorities do not always align with individual or societal interests. Regulation is increasingly important as generative information access becomes more integrated. This will involve internal company policies and external laws such as GDPR or future AI regulations. Unlike traditional search engines, where removing specific content can address privacy concerns, generative models create new challenges because there is currently no established way to trace or erase learned content without retraining an entire model. Future research should focus on understanding how architectural decisions affect the capabilities of generative systems while ensuring that these advances serve technical progress, user needs, ethical standards, and regulatory requirements.

3.6 Broadening the IR Community

Advances in AI, especially in ML, deep learning, and NLP, are central to IR research. In the next five years, ongoing collaboration and shared challenges between these fields will strengthen their integration. As a result, IR research will become more prominent at major AI and NLP conferences. IR has unique characteristics and challenges that could lead to methodological breakthroughs with broad implications for other research fields. IR researchers bring experience in representing documents and large collections for effective information access. This background gives the IR community unique strengths that complement ML and NLP. Expertise developed through work with diverse document types, languages, modalities, and established evaluation benchmarks will be valuable for addressing challenges in RAG and developing robust domain-specific or multitask models for personal and professional applications.

Additionally, experience in IR in interpreting explicit relevance labels and implicit behavioural signals is vital for instruction tuning and reinforcement learning, offering significant knowledge transfer potential. The community's understanding of achieving run-time and storage efficiency is crucial for developing larger models and distilled smaller models, impacting scientific advancements, cost reduction, and democratising AI with long-term climate and societal effects. Furthermore, IR expertise in evaluation, especially user-centric aspects, can significantly influence research on isolated tasks in ML and NLP, such as question-answering and summarisation, by contributing to robust evaluation methods that detect super alignment and separate it from evaluation artifacts.

3.7 Obstacles and Risks

The architectural redesign of generative information access demands extensive training, fine-tuning, and evaluation data, which is challenging to acquire outside the industry. While simulation and LLM-labeled data provide alternatives, they may compromise scientific rigour. The IR community still faces the challenge of obtaining high-quality, real-world data for reinforcement learning and instruction tuning, emphasising the importance of public and shareable evaluation data. The current scale of information access tasks surpasses traditional methods, necessitating a concentrated effort on shared interests within the IR community.

Resource constraints are significant, as training LLMs and even fine-tuning existing ones require substantial computational resources, often inaccessible in academia. The IR community must find ways to share resources and ensure that access to the largest AI models is not essential for research. Additionally, the rising costs of ML and LLM research, including infrastructure and expertise, demand increased funding and collaboration between industry and academia. This collaboration is vital for supporting open science practices and training future industry professionals. Furthermore, expanding the IR community globally is crucial to harness diverse perspectives and ensure a skilled workforce, focusing on diversity, equity, and inclusivity to shape the future of generative information access.

4 What is IR in New User Experience, Information Access & Interaction Scenarios?

4.1 Description and Motivation

Understanding user engagement with information systems is essential for developing effective IR solutions. Researchers across disciplines have long studied how users define and pursue their information needs. The introduction of LLMs and GenAI rapidly changes these interactions. Users now hold dynamic, interactive conversations with AI systems that can access diverse information sources, which alters traditional IR tasks and evaluation methods. Researchers must study user behaviour in modern settings to keep pace with these changes. Conversational interfaces and features like ChatGPT's "canvas" mode let users interact directly with documents instead of relying solely on prompts. As interaction styles become more complex, examining how users engage in these scenarios is crucial to designing IR systems that better support the discovery and use of information.

4.2 Research Challenges and Proposed Research

4.2.1 What is Information and What is IR

GenAI introduces a fundamentally different approach than traditional IR retrieval systems, offering new opportunities while posing questions and challenges. A key question is understanding what constitutes "information" within an LLM. While LLMs memorise and encode world knowledge, this knowledge differs from traditional forms, such as database tuples or knowledge bases. This raises issues of the definition of IR today, user interaction with information, and the implications of not fully understanding how LLMs create responses from stored information.

This section outlines several research directions to address these challenges. First, it is crucial to comprehend how knowledge is encoded in LLMs, its decomposition, reconstruction, and reliability, as well as internal knowledge and user-facing content. Understanding how this dispersed knowledge relates to more traditional data storage and its implications for users is essential. Additionally, exploring how users perceive and interact with the generated content, which appears to be authored yet is ephemeral, is also important. Understanding each step in an inferred output for a user, understanding the user's mental models of these systems, and considering the increasingly blurred boundaries between modalities and system types are also key focus areas. Finally, these considerations impact evaluation methodologies and the evolving definition of IR.

4.2.2 Keeping Up

GenAI tools are being deployed in many different settings. Recently deployed tools enable users to: *(i)* engage in dialogue; *(ii)* summarise a document; *(iii)* ask questions about a specific document or set of documents; *(iv)* ask for writing assistance; or *(v)* ask questions about an image. As GenAI technology improves, novel tools will likely be released unprecedentedly. This is a unique opportunity for the IR research community. IR researchers will need to study and understand users in the new information interaction scenarios that are being unleashed rapidly. Each scenario may vary along different dimensions: *(i)* the user's intent; *(ii)* how the request is posed to the GenAI system; *(iii)* how the GenAI output is incorporated into workflows; *(iv)* the implicit signals that may predict whether the GenAI output was satisfactory; or *(v)* how a user may reformulate a request; and *(vi)* the factors that may influence post-task satisfaction. Perhaps most importantly, we need to understand the long-term impact of such tools. Are GenAI tools that provide writing support, making people better writers? The answer may not be straightforward. It may depend on the way that the assistance is provided. Finally, are specific tools being repurposed to fulfil interesting needs? For example, are LLM-based chatbots being used for second-language learning?

4.2.3 How Tasks Impact User Expectations

Integrating LLMs and GenAI into user information-seeking processes transforms how users interact with information systems, enabling new methods of information delivery and presentation. These advancements allow AI and IR systems to facilitate traditionally human-to-human interactions, such as providing overviews or clarifying concepts, thereby lowering barriers for users to address their information needs across various task contexts. As these technologies evolve, it is crucial to reassess how user tasks influence their goals and expectations for interaction and IR.

Several areas warrant investigation: the emergence of new types of information needs, including aspects of criticality, temporality, and context; the types of tasks users want to be supported by next-gen IR/GenAI systems and the appropriateness of conversational approaches; and how expectations and behaviours are shaped by the criticality of the information sought. Additionally, different task types impact the desired interactions, such as when users prefer a list of documents versus a direct answer or how to support users in critical decision-making, learning about new topics, or accomplishing complex tasks. For instance, a doctor may require a structured list of treatment options with justifications, while a user learning a new topic may seek interactive guidance and proactive system support. Understanding these dynamics is essential for optimising AI-assisted information interactions.

4.2.4 Interaction

Recent technological advances have transformed user-system interactions, particularly in IR systems. Traditionally, users played an active role, adapting their information needs into keyword-based queries, while systems were reactive. Feedback was limited to qualitative signals attached to single-answer items. However, with LLM-based systems, users can interact using natural language, providing more nuanced feedback and engaging in conversational search. These systems support session context and mixed-initiative interactions, where the system can actively participate by asking clarifying questions. This shift alters user mental models of information access, leading to new expectations and behaviours in interactive IR systems. Despite efforts to incorporate user feedback without new queries, such as modifying document lists, these approaches have not aligned with users' mental models. Research has focused on encouraging explicit feedback and analysing interaction signals like clicks and dwell time. The engagement with conversational systems raises questions about interpreting user feedback through natural language, expected responses, and the role of relational elements (e.g., apologies).

As natural language interactions become more prevalent, systems adapt by providing feedback through direct responses, query suggestions, and result diversification. This evolution raises questions about the effectiveness of natural language in specifying and refining information needs, alternative methods for presenting refinement options, and how diverse alternatives should be presented. Users attribute agency to systems in this new interaction model, creating opportunities for mixed-initiative interactions. This transformation raises questions about how mixed-initiative, clarifying questions, and proactive interactions affect user experience and behaviour. The concepts of implicit and explicit feedback must be reconsidered in the context of GenAI IR systems, where traditional signals like queries and clicks may change. There are opportunities to assess user satisfaction and understanding through GenAI and IR systems feedback. Research questions include applying human-human communication models to human-AI conversations, understanding detailed user feedback, and exploring new implicit feedback signals. These signals could include interaction patterns, engagement with other devices, and biometrics data, offering insights into user satisfaction and understanding.

4.2.5 Ephemeral Documents

IR traditionally relies on implicit feedback from user interactions with documents to understand user preferences and evaluate system quality. This involves users submitting queries to search engines and interacting with ranked documents, with metrics like click-through rate and user engagement assessing system performance. However, in the era of GenAI, where traditional documents are absent, new challenges emerge for user and algorithmic studies. Algorithms must adopt novel approaches to learn from user interactions, as the lack of collective interaction data limits understanding of system failures and the complexity of tasks or queries for users.

The ephemeral nature of GenAI responses introduces additional user interaction challenges. Unlike traditional IR, where users can revisit documents over time, generative systems like ChatGPT simulate this by storing user interaction history. This raises concerns about privacy and market fairness, as users are restricted to accessing past interactions within the same system. Future research directions include defining what constitutes an "information unit" in the GenAI era and how to store them for user access, supporting users in revisiting previous interactions, understanding novel interactions with old dialogues, and addressing the implications of evolving systems that provide new information. Additionally, it is crucial to explore user expectations regarding compatibility with previous content and interactions across different GenAI systems.

4.2.6 Response Credibility

In the traditional 10-blue-link search engine results page (SERP) setup, users rely on various information channels, such as search snippets and page URLs, to assess content credibility. Over time, users develop preferences and trust specific news sources for reliability. The quality of written content often indicates the writer's trustworthiness, influencing users' decisions to trust the information. For instance, users may prefer genuine buyers over potentially biased sponsored reviews for product reviews. This approach allows users to form well-rounded opinions by reading multiple reviews rather than relying on a single AI-generated summary. However, interacting with AI-generated content can obscure these experiences, as the uniformity of AI writing may not reflect the content's credibility.

GenAI system users face challenges in evaluating the credibility of generated content, as they might mistakenly believe that the AI system is infallible. This necessitates AI systems to be transparent about their confidence levels and inaccuracy potentials. Currently, RAG systems cite sources similarly to scientific articles, using numbered citations and a reference section. While intuitive for system designers, this method may not be user-friendly, leading to reference-checking difficulties. To enhance response credibility, research should explore how citation styles affect user perceptions, the ability to distinguish credible content, and ways AI systems can provide helpful explanations. Additionally, novel presentation methods for RAG content, such as embedding it in SERP-like user interfaces, could improve user experience and content consumption.

4.2.7 Conveying Confidence

Traditional search engines convey confidence through ranking, giving users a signal of the system's confidence in the results, along with URLs, page titles, and query-biased text snippets for user assessment. In contrast, GenAI systems output synthesised text, sometimes with citations, but often lack clear indications of confidence levels. Confidence can vary for components, such as individual "information nuggets" and logical reasoning across concepts. Important research questions include how LLMs should convey their confidence in answers and factual statements, indicate when they are making inferences, and communicate the extent of these inferences. Additionally, it is crucial to explore how users can be given control over the confidence level required for different tasks.

4.2.8 Self-Regulated Learning

Self-regulated learning is a process where learners monitor and control their learning through understanding tasks, setting goals, employing strategies, monitoring progress, adapting, and reflecting on outcomes. Effective self-regulated learning engagement enhances learning outcomes. GenAI technologies can support self-regulated learning by aiding task understanding, providing goal feedback, encouraging strategic learning, offering feedback on notes, and promoting self-reflection after learning tasks.

4.2.9 Learning by Doing

The search-as-learning research community focuses on understanding how people learn through search engines and AI systems, emphasising that better learning outcomes are associated with less time searching and more time engaging with content. Traditional IR research has been limited to helping users find information, but future research should explore new ways for users to interact with the content. GenAI technologies offer promising opportunities to enhance document-level interactions by creating tools that test understanding, generate questions, summarise content, and connect documents. These tools can help users resolve information needs directly within the document, rather than returning to the search interface, by linking documents that corroborate, contradict, provide evidence, elaborate, or describe prerequisite and follow-up concepts.

4.2.10 The Role of Serendipity in Search

Serendipity is a fundamentally important concept in search, where users often discover unexpected yet helpful information, aiding in uncovering "unknown unknowns" and enhancing search outcomes. However, with the advent of GenAI technologies, which provide precise answers, the likelihood of serendipitous discoveries may diminish, potentially leading to shorter, less engaging search sessions and overestimating knowledge breadth. Future IR research should explore how GenAI can support serendipitous discovery by recommending follow-up questions, showing related concepts, and linking generated responses to documents that offer supporting evidence and related information.

4.2.11 Future AI+IR to Support Users

Integrating GenAI and other emerging technologies into the information-seeking process offers vast potential for enhancing user-system interactions by effectively understanding and modelling users. This involves considering their situation, context, cognitive state, and personal traits, as well as supporting various types of information work such as locating, assessing, analysing, and decision-making. These advancements are particularly beneficial for complex tasks requiring multi-turn, mixed-initiative conversations between users and systems. This section explores several ideas and examples of how GenAI, combined with IR technologies, can support users in these endeavours.

A key area for research is understanding the various ways GenAI and IR systems can support users. These systems may serve as information providers, tutors, advisors, or collaborative partners. Important questions include which roles users prefer in different situations, what capabilities or limitations are associated with each role, and whether users want interactions that resemble human conversation or favour more direct, efficient exchanges with AI. Another important direction involves exploring how multimodal GenAI and IR can enhance user support by combining text, images, audio, or video. Multimodal content has the potential to make complex information easier to understand. Researchers should investigate when it is most effective to present multimodal content, how much control users should have over selecting content types, and how these approaches impact user comprehension and engagement.

GenAI and IR technologies can improve accessibility for users with a range of impairments by making it easier to seek out and use information. Key research directions include adapting how information is presented to different user groups and assisting users in expressing their information needs, even when their prompts are incomplete or unclear. Another promising area involves integrating GenAI with documents so that the system can act as a personalised tutor, adjusting explanations based on each user's level of knowledge. This approach raises important questions about balancing ease of learning with true understanding, how personalisation uses context, and how much control users should have over what the system learns about them.

Integrating AI and IR technologies directly into users' existing tools, such as email clients or word processors, can make these systems more useful without disrupting established workflows. Instead of forcing users to switch between separate applications, this approach embeds intelligent features where needed most. Key research questions include how to add AI and IR capabilities to support productivity without causing distractions and how much control users should have over when and how these features are activated. This direction supports a shift toward seamless, user-friendly information access rather than interrupting users' work with separate retrieval tasks.

4.3 Broader Impact: Influence on Other Research Fields and Society

The exploration of "what is information inside an LLM" is crucial as it influences fields like NLP and recommender systems, which must also address this question and its user implications. As the roles of search, recommendation, and conversation increasingly overlap, these disciplines must address similar questions about user interaction and system design. Studying mixed-initiative approaches and feedback mechanisms in NLP and recommender systems can inform new methods in IR. Adapting these insights will help build stronger foundations for future IR research while improving how all three types of systems meet users' information needs.

4.4 Broadening the IR Community

An updated definition of IR would enhance the IR research community's collaboration with other fields, especially as users' expectations for AI and IR systems grow. To meet these expectations, the IR community must expand its expertise by integrating insights from education, learning, communication, and journalism to present information effectively. This involves structuring information summaries, scaffolding learning, generating user-appropriate text, and more pervasive information presentation. Additionally, developing new metrics based on cognitive psychology, decision-making, and learning theories is crucial for assessing user satisfaction and the effectiveness of AI and IR systems. Aligning IR systems and evaluation methodologies with user information needs and understanding diverse user behaviours, preferences, and values will support the development of new tasks and evaluation methodologies, ensuring representation in interactive evaluation approaches.

4.5 Obstacles and Risks

The question of what constitutes information within an LLM remains elusive, with potential answers being limited or applicable only to smaller models. This uncertainty may restrict the anticipated benefits of understanding LLMs. The rapid evolution of information access and retrieval tools poses challenges for interaction research, making it difficult to study various modalities in depth and to achieve a comprehensive understanding of user interactions, unlike the more stable environment of traditional search engines.

A significant risk is the lack of collaboration between researchers and companies leading the development of new information tools, which may result in missing valuable user experience insights and the establishment of sub-standard systems. This is evident in the standardisation of interfaces in chatbot LLM systems. Additionally, technological advancements can create barriers or enhance accessibility for marginalised communities. To address this, diverse user involvement and collaboration with accessibility researchers are crucial. Furthermore, exploring methods to mitigate risks such as bias, hallucination, and privacy concerns in GenAI systems is essential.

5 Agentic Information Retrieval (AIR): IR for All

5.1 Description and Motivation

IR has traditionally focused on processing queries, but there is a growing recognition of the need to shift towards task-based IR. This approach emphasises understanding and addressing the broader tasks users aim to accomplish rather than just responding to isolated queries, see Figure 1. For instance, instead of issuing multiple queries about solar panels, users could input their overall goal into an IR system, which would then provide a comprehensive research report, recommendations, and an action plan. This task-oriented approach can be applied to various scenarios, such as planning vacations, scheduling events, conducting research, and making decisions.

Task-based IR systems can be enhanced by employing agents that understand the context and nuances of the user task, offering more relevant and comprehensive results. These agents can operate in three modes: Assistant, Collaborator, and Mentor. In the *Assistant* mode, the agent acts as a helper, providing timely information and handling straightforward tasks like booking travel tickets. The *Collaborator* mode allows the agent to take initiative and make recommendations, such as helping a parent design a study plan for their child. The *Mentor* mode involves the agent

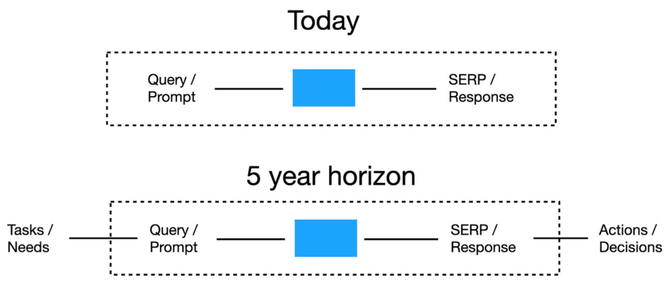


Figure 1. A vision of IR over the next 5 years as we expand the input beyond queries/prompts and the outputs beyond responses to actions and decisions.

acting as an expert, guiding users through complex tasks, like advising an academic researcher on starting a new hobby such as golf.

Integrating these modes enables users to engage with IR systems in a way that best suits their needs and context, focusing on higher-level goals. Extensive research into user behaviour, task dynamics, and context is necessary to realise this vision, alongside developing capable agentic systems and new evaluation metrics. Additionally, it is crucial to consider the societal implications of these advancements in IR technology.

5.2 Proposed Research

5.2.1 Understanding the Task

The transition from systems that reply to queries to those that address tasks creates several research questions. Researchers and engineers need to establish a common terminology and hierarchy for discussing tasks, which could benefit from a literature review, including studies outside of IR that explore tasks in daily life or work. User studies could identify areas where people desire agent assistance. Representing tasks in a system-usable format is crucial, with past suggestions including Markov decision processes and graph models, and these could be enhanced with semantic task descriptions and user models. Collaboration with planning and dialogue research communities could be beneficial. Another area of interest is recognising when someone is engaged in a task based on interactions, such as search or system interactions. Simple techniques like clustering have been effective, and ML models could further improve task recognition, especially with a limited set of tasks.

Once a task is recognised, the system should be able to decompose it into achievable sub-tasks, compose tasks to understand the bigger picture and suggest best practices to navigate complex tasks, see Figure 2. Techniques from query suggestion might be applicable, but new work in sequence prediction may be necessary. Engaging with the searcher to refine task understanding is also important. This could involve using interaction data, conducting user studies, or leveraging existing test beds like TREC to map task titles to comprehensive descriptions. Multiple methods could help develop a vague task understanding into a precise description that an agent can execute.

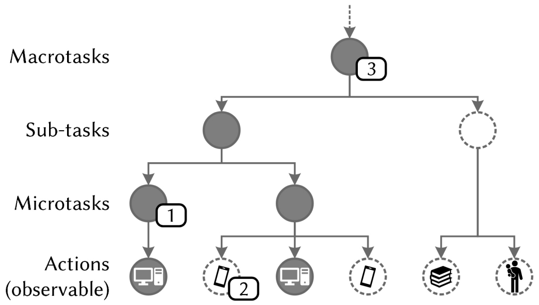


Figure 2. A hierarchical model of task understanding. Figure from Shah, Chirag, Ryen White, Paul Thomas, Bhaskar Mitra, Shawon Sarkar, and Nicholas Belkin. "Taking search to task." CHIIR. 2023.

5.2.2 Understanding the Context

Effective agentic IR systems must integrate a hierarchy of contextual dimensions to align task execution with user needs. These dimensions include real-time spatiotemporal context, user-personalised context, social context, task-sequence context, and negative context. Real-time context involves temporal and geolocation factors, while user-personalised context requires dynamic modelling of intrinsic factors like intent and decision-making patterns. Social context considers collaborative aspects and task-sequence context models dependencies between tasks. Negative context involves identifying misleading information to prevent errors. These dimensions create a complex data landscape, necessitating robust frameworks to resolve conflicts and prioritise contextual signals.

Contextual data is acquired through sensor-driven real-time input and historically derived personalisation. Real-time context is captured via device sensors and environmental APIs, while historical personalisation uses longitudinal user data to train preference models. For new tasks, cross-task transfer learning addresses cold-start issues. Challenges include maintaining temporal consistency and protecting privacy. Personalised context in task-oriented IR systems involves a three-stage pipeline: retrieval of relevant data, filtering based on user preferences, and context-aware ranking to optimise utility. This process requires multi-objective optimisation frameworks with dynamic weighting mechanisms.

5.2.3 Understanding the User

Personalisation in agentic IR requires a comprehensive understanding of user behaviour and context at both macro and micro levels. At the macro level, it involves modelling user decision-making through behavioural patterns and factors like intrinsic needs, social influences, and environmental constraints while also considering the reciprocal influence between users and agents. At the micro level, personalisation focuses on intent disambiguation and preference alignment, using explicit and implicit signals to tailor information presentation and maintain situational awareness. Challenges include avoiding over-personalisation biases and ensuring transparency, especially in critical areas like healthcare and finance. Future research should integrate macro-level behavioural theories with micro-level interaction analytics for a holistic and ethical approach to user understanding in agentic IR systems.

5.2.4 Supporting Decisions and Actions

Supporting complex tasks involves providing information and facilitating higher-order activities such as decision-making and action-taking. Decisions require synthesising information from multiple sources, and systems can assist by offering action plans that need validation for feasibility, efficiency, and effectiveness. Depending on user preferences, systems can perform actions on behalf of users, necessitating an understanding of tasks and system capabilities, including available agents or tools for task execution. Systems must decompose complex tasks into subtasks, formulate action plans, and determine which tasks are best handled by users or agents. They should use resources effectively for decision-making and action, clarify the role of search engine responses, and employ algorithms and data stores to map tasks to action plans or agents. Matching algorithms are essential for finding appropriate tools and agents, and agent architectures require protocols for communication and accessing tools. Specialised agents with specific expertise or characteristics, such as on-device or offline agents, are also needed.

The user experience should empower users to decide which subtasks to perform or delegate, providing necessary information for decision-making and allowing comparison of outcomes. It should explain system actions, consider factors like robustness and privacy, and be aware of other agents and tools, such as OpenAI's Operator agent. Users should be able to revisit past actions to understand decision rationales, building trust and confidence in the system. As agents progress in tasks, they should share updates, receive real-time feedback, and raise exceptions with users. They should also reveal task completion processes to help users learn and understand tasks. This approach supports user adoption and willingness to use agents for new tasks.

5.2.5 Develop Datasets and Benchmarks for Evaluation

To ensure agents' success in their respective tasks, evaluating them based on their specific applications and various dimensions of effectiveness is crucial. This evaluation should encompass agent-user interactions and the success of the agent's specific subtasks. Existing IR metrics alone are insufficient for this purpose, necessitating the development of new datasets, frameworks, and evaluation metrics. Three key research streams have been identified for this endeavour. Firstly, establishing a ground truth for task execution success involves defining objective criteria to measure the agent's performance accuracy and effectiveness. This includes creating comprehensive tasks with predefined outcomes and employing precision, recall, and F1 score metrics. Real-world scenarios and user interactions should be incorporated to ensure practical relevance. Secondly, developing benchmarks for personalisation requires identifying user behaviour and preferences to enhance user experience. This involves analysing user interactions, search history, and feedback to create personalised models, with benchmarks including user satisfaction and task completion rates. Lastly, user studies are essential for understanding the usefulness and satisfaction of IR agents. These studies involve experiments capturing qualitative and quantitative data on user experiences, using metrics like Net Promoter Score, System Usability Scale, and task success rates to evaluate user satisfaction and identify areas for improvement.

5.2.6 Personalisation

Integrating Agentic IR with personalisation requires careful consideration of sensitivity and levels of personalisation. Sensitivity involves determining which tasks need personalisation and when appropriate, as not all tasks, such as booking doctor appointments, prioritise personalisation. Additionally, personalisation must be managed to prevent unintended privacy leaks, such as booking a specialist during a group meeting. Levels of personalisation focus on achieving high satisfaction for individuals while balancing the extent of personalisation needed and potential concerns. For instance, while detailed actions can enhance personalisation, there may be discomfort in allowing an agent to comprehensively analyse personal relationships to suggest suitable gifts.

5.3 Research Challenges

5.3.1 Task Modelling

Understanding and decomposing tasks into individual information needs is crucial for the success of an IR system, requiring both immediate intent and overall goal comprehension. Key challenges in task modelling include representing a task in a system-usable format, such as a tree of subgoals, actions, or sequences in a graph or high-dimensional space, which may require explicit or implicit user feedback. Recognising a task in progress and understanding the intent is essential to determine the appropriate time for support. Composing and decomposing tasks into sub-tasks and actionable steps involves predicting or recommending steps based on task knowledge and observable actions. Finally, exposing the reasoning and decomposition to the searcher is vital to demonstrate task structures, provide control over agent assistance, and determine the appropriate level of detail and language for feedback and control.

5.3.2 Personalisation and Shadowing the Person and Learning

Agentic IR often requires systems to personalise actions for specific tasks and individual users. Achieving effective personalisation depends on capturing user preferences and behaviours while maintaining privacy. One promising approach is for agents to observe how humans perform tasks, allowing them to learn new actions through shadowing rather than relying on explicit task templates. Personalisation may also involve coordination among multiple specialised agents, enabling more adaptive and expert support for each user's needs.

5.3.3 Privacy, Safety & Reliability

As the agent is personalised, there could be various potential safety issues. Agents with sensitive personal data should have robust authentication mechanisms to ensure that malicious actors cannot access them. These include other users as well as agents that communicate with each other. The Agentic IR systems should facilitate fine-grained control over which personal data streams agents could access and how they can be shared with other entities. As agents operate longitudinally, they could exhibit biases in the outcomes. Eventually, the biases could impact users without them even noticing. Research challenges lie in how to evaluate and prevent this.

5.3.4 Trustworthiness, Transparency & Explainability

As agents execute actions on behalf of users, user trust towards agents is paramount for the sustainability of Agentic IR. Transparency and explainability can enhance the trust between users and agents. While current research on explainability often focuses on explaining atomic decisions, agents will need capabilities to provide personalised explanations that often involve complex information and a series of decisions/actions.

5.3.5 Interaction and Feedback

Agentic agents will have to measure and utilise implicit and explicit feedback from people. This will include both signals for task success: how to detect these implicitly and when to ask the person about this explicitly. It is also more likely that we will have less explicit feedback, such as clicks, but more natural language interactions – which will also create challenges in interpretation and scalability. Then, interaction mechanisms are needed that involve proactive interaction with the person when the agent requires more information; for example, asking clarification questions. The agent will also require user input to determine which subtasks it could execute, when to prompt the user for decisions, the level of control, and access permissions provided to the agent. Finding optimal mechanisms to proactively prompt the user for such information without hindering the user experience is an unsolved research challenge.

5.3.6 Evaluation

As there is no clear ground truth to evaluate the success, measuring whether/how well the agents' actions satisfy user preferences could be very challenging. For instance, if the agent is tasked with booking an air ticket, user satisfaction will depend on cost, arrival and departure times, or even the specific seat selection. While the outcome is important, other attributes, like the justifications provided by agents and the way agents present information, should also be considered in the evaluation. Furthermore, since evaluation could be highly task-specific, developing, benchmarking and validating generalisable Agentic IR systems can be resource-intensive. Defining the completeness of a task and figuring out the appropriate time to conduct the evaluation is another research challenge while considering the task's time and context.

5.4 Broader Impact: Influence on Other Research Fields and Society

To realise the vision depicted in Figure 1, collaboration with various communities is essential, particularly for addressing the right-hand side of the diagram, which involves task execution and validation in real-world scenarios. This includes tackling issues of trustworthiness, robustness, and transparency. The IR community can contribute significantly by providing insights into the task and user understanding, modelling, and evaluation, along with tools for managing information. Key areas for collaboration include addressing accumulated biases, as agents may inadvertently reinforce existing biases in data, necessitating partnerships with ethics and fairness researchers. *Over-personalisation* can lead to echo chambers, emphasising the need for strategies to ensure diverse information exposure. Agents can aid in predicting and responding to large-scale unexpected events, such as natural disasters or pandemics, by analysing data in real-time, requiring input from crisis management and public health experts. The societal impact of personal agents calls for studies on their social influence and ethical implications. Developing advanced dialogue systems and task orchestration capabilities also involves integrating NLP and workflow management, benefiting from advancements in these fields. Finally, ensuring agents operate within legal and regulatory frameworks is crucial, necessitating collaboration with legal experts to address data privacy, security, and compliance issues.

5.5 Broadening the IR Community

Agentic IR envisions a comprehensive system that aids in task completion and decision-making, integrating a range of technologies beyond traditional IR. Task execution might be managed by technologies like workflow and execution engines, expanding the IR community's role in understanding and contributing to these areas while maintaining its core mission of facilitating task completion. This research direction will broaden the IR community by fostering interdisciplinary collaboration, as developing intelligent agents will require expertise from fields such as AI, HCI, and cognitive science. This collaboration will enrich the community with diverse perspectives and methodologies. Additionally, focusing on task-oriented agents will enhance user experience by creating intuitive systems that meet users' needs, shifting from query-based to task-based interactions and attracting a wider audience. The development of agents also opens new research opportunities in personalisation, context awareness, and real-time decision-making, encouraging innovation and contributions from more researchers. Furthermore, agents have practical applications in scenarios like virtual assistants and automated customer service, demonstrating the practical benefits of IR research and increasing the community's visibility and impact, attracting more funding and industry interest. In summary, agentic IR will deepen the roots of IR while opening it to more stakeholders and disciplines, renewing interest and investment in the field.

5.6 Obstacles and Risks

The implications of highly effective agents taking over tasks could lead to users becoming less skilled or unaware of how to perform them. This highlights the impact of local legislation, such as the European AI Act and Digital Services Act, which can restrict data usage for training, prohibit model sharing, and affect system design due to transparency requirements. Additionally, the importance of accountability when agents perform high-level tasks such as making purchases or booking appointments, as mistakes could have serious consequences. Future research should focus on developing safety mechanisms and regulatory guardrails to protect users from potential scams and errors when using Agentic IR systems.

6 LLM-based Simulation for Evaluation

6.1 Description

Offline evaluation of information access systems is evolving with LLM-based approaches, which promise to redefine traditional methodologies like Cranfield and fixed recommender-system datasets like MovieLens and Netflix. This section discusses LLM-based simulation for evaluation, where "simulation" refers to AI-generated content that closely mimics human-generated content. All aspects of offline evaluation, including relevance judgements, queries, clicks, reformulations, metrics, and documents can be simulated with LLMs. We explore this development's potential benefits and risks, the methods for validating simulations, and the broader implications to other closely related research areas.

6.2 Motivation

Testing new components in information access systems often involves offline evaluation, which uses datasets of content with labels or recorded user interactions. This method simulates human interaction with the system, such as in IR tests where queries, relevance judgements (QRELs), and evaluation measures mimic initial searches. Offline evaluation is cost-effective and repeatable, but creating datasets is expensive and time-consuming, often requiring extensive manual effort. The labels may reflect a single viewpoint, and specific evaluations, like user feedback, are challenging to incorporate. Crowdsourcing can be used, but remains costly and slow, with challenges in generalising results due to the underrepresentation of minorities.

Researchers recognise the limitations of offline evaluation and have explored extensions, particularly with the advent of LLMs, which can simulate human relevance judgements. This suggests the potential for simulating other aspects of information access, making evaluation richer and more dynamic. Simulations allow precise control over evaluation parameters, aiding hypothesis testing and understanding cause-effect relationships. However, there are risks, such as the need for reevaluating statistical comparisons and ensuring validation to avoid biased results. Validating LLM simulations is a new research area, focusing on accurately simulating properties to ensure metrics on simulated collections reflect real experiences.

6.3 Proposed Research

Simulation of Test Collections and their Components. Two main aspects of test collections where LLMs can be used, assessments and content, are considered here. In terms of assessments, LLMs can address the incompleteness of ground truth by acting as digital twins of original assessors, thus continuing their work. They can also simulate assessor disagreement by behaving differently yet plausibly, offering insights into system behaviour under varying judgements. LLMs can emulate different personas, reflecting diverse backgrounds and biases, to better understand how systems serve varied users. They can transform traditional judgements into more nuanced forms, such as graded or preference judgements, and handle private or harmful content, allowing evaluations without human exposure. Beyond topical relevance, LLMs can assess other dimensions like correctness or conciseness and evaluate content scope beyond single documents, providing insights on missing content or redundancy. Regarding content, LLMs can enhance topic descriptions by introducing variations or clarifications in narratives and relevance criteria. They can generate query variations by adopting different user profiles and creating comprehensive queries reflecting diverse user backgrounds. LLMs can also generate new topics with corresponding narratives and relevance criteria, and produce new documents to improve coverage or recall for challenging topics. This content generation can be conditioned on specific topics, enhancing relevance as a byproduct.

Simulated User Interactions. LLM-based AI agents can simulate user behaviour in IR studies, traditionally involving human subjects in lab settings. These agents can mimic user profiles and demographics, performing tasks and providing interaction feedback, thus enabling large-scale user satisfaction and engagement measurements. This approach allows for creating extensive interaction datasets, helping identify IR system design issues and suitability for diverse user groups, including minorities and people with disabilities. Beyond single-user studies, LLM simulations can generate large-scale conversation datasets, evaluating system performance for varied users. LLM-simulated web users can audit commercial IR systems, revealing how recommendation algorithms function for different user types.

Evaluation Metrics and Methodology. Using LLMs for relevance assessments necessitates novel aggregation methods for relevance values, preserving distributions rather than averaging. Simulated users may lead to new evaluation approaches, focusing on user satisfaction instead of traditional metrics. LLMs could also serve as metrics, evaluating system outputs numerically. The ability to handle preference judgements with LLMs opens possibilities for new metrics.

Counterfactual Scenarios and Simulation Processes. LLM-based simulations enable the exploration of counterfactual scenarios in test collections and user studies, examining potential user actions and their outcomes. Synthetic data generation involves modelling unobserved elements like user information needs, with processes conditioned on existing data. The challenge lies in identifying optimal simulation processes for trustworthy data. Parameterised simulations, controlling aspects like document length and language complexity, offer insights into system effectiveness across dimensions, enhancing understanding of system behaviour.

Evaluating, Validating, and Trusting Simulations. The validation of simulated evaluations is analogous to human evaluations, focusing on two main aspects: experimental outcomes and human behaviour data. For experimental outcomes, the goal is for simulations to produce results consistent with human-based experiments. For instance, if a simulated evaluation using synthetic data finds that ranker B is better than ranker A, just as a human evaluation does, the simulation is considered valid. However, differences in sensitivity between human and synthetic evaluations can occur, potentially leading to biases. For example, if a language model-based relevance judge is biased by specific sentence structures, it might incorrectly favour one system. This highlights the need for comprehensive human experiments to validate synthetic evaluations effectively. When comparing human-generated data to synthetic data, agreement measures such as Krippendorff's alpha are valuable for assessing the quality of synthetic annotations. High agreement with human annotations indicates more reliable synthetic annotators. However, it is also essential that generated artefacts, such as queries and documents, not only resemble plausible human-created items but also capture the diversity found in real-world data. Establishing standard evaluation methods and identifying potential biases or errors in synthetic data are important steps towards this goal.

A key challenge lies in understanding how humans approach these tasks and designing tests that can accurately compare human and synthetic outputs, which may require hybrid or independent human evaluation processes. This is particularly relevant when using LLMs, as their training data may not include certain types of knowledge, such as oral traditions or lived experiences, that influence authentic human responses. To ensure meaningful simulation of diverse perspectives, extensive validation is necessary to determine whether LLMs can effectively represent responses from various personas, for example, a “Black trans woman living in an Atlantic metropolitan city in the United States”. Further research is needed to map out which personas can be realistically simulated and identify these simulations’ boundaries. LLM-based simulations should be validated across several dimensions, including demographics (age, gender, location, language), domain-specific knowledge (law, medicine), user experience (novice vs. experienced users), socio-economic status, geographic location, and disabilities. This comprehensive approach will help ensure that simulations accurately reflect the diversity of human experiences and interactions.

The Impact of Simulation on Evaluation Methodology. Research on uncertainty in estimating system performance using Cranfield experiments highlight the impact of variability from document collections, queries, relevance assessors, and pooled systems on the stability of system ordering by metrics. Traditionally, variations were generated through bootstrapping, collection properties, and new user queries, with crowdsourced workers providing additional relevance assessments. These methods demonstrated that such variations could influence system performance and ordering. With LLMs, there is potential to generate variations for each source of uncertainty, raising questions about modelling system performance over multiple dimensions, comparing systems statistically, and determining the feasibility of running systems over a multiverse of combinations.

Challenges include how to treat LLMs, whether to invoke them multiple times to represent different users or prompt them for unique queries for the same information need. Issues like assessor disagreement introduce uncertainty in system performance estimates, and modelling errors from crowdsourced workers involve assumptions about discrimination and bias. By generating multiple components of an experiment, researchers can report variations in system performance, offering predictions for future experiments with evolving collections and queries.

6.4 Research Challenges

The key issues focus on how using LLMs to simulate elements within the evaluation pipeline can influence utility and impact. One consideration is identifying which components are most beneficial to simulate and determine the types of problems or scenarios where simulation offers clear advantages. Another crucial aspect involves evaluating and validating these simulations to ensure they generate data and outcomes that closely reflect real-world conditions. This includes establishing reliable features and methodologies to prevent misleading results, as well as developing robust meta-evaluation practices and clear guidelines for good practice. Finally, simulation has the potential to change evaluation methodology fundamentally; introducing new data types could make traditional approaches obsolete, prompting a rethinking of how results are analysed, reported, and measured, including possibilities such as directly assessing user satisfaction.

6.5 Broader Impact: Influence on Other Research Fields and Society

We see two areas of research influence: *(i)* computer science colleagues in research areas related to SWIRL and *(ii)* social sciences such as library science, legal, and education. Evaluation methods in NLP and ML are shifting from traditional metrics like ROUGE and BLEU toward using LLMs to judge outputs. This is especially evident in Reinforcement Learning from Human Feedback and Direct Preference Optimisation techniques. Simulating diverse user profiles for automated evaluation offers opportunities to enhance these approaches. To address the demand for better data and evaluation metrics, developing information access corpora and public leaderboards can help broaden participation in research and lower barriers to entry. As LLMs become more capable of generating synthetic data and simulating users, there may be less reliance on crowdsourcing for data generation, a trend that could have economic effects on communities currently dependent on these tasks. However, simulated data can improve quality while reducing costs, particularly if validation mechanisms are used to ensure trustworthiness. In addition, bias remains a concern with LLM-generated content. By focusing on detecting and mitigating biases during simulation, these technologies can contribute to fairer AI systems. In fields like library science and legal discovery, simulated users or automated relevance judgements could support better access to information and maximise recall across entire document collections. Education also benefits; LLMs can generate student submissions that meet varying rubric criteria to help train teachers in assessment practices. They also offer new ways to test plagiarism detection tools using controlled synthetic examples. Furthermore, automating content moderation through AI-based evaluations may protect human moderators by reducing their exposure to harmful material. Simulating various user backgrounds also allows researchers to improve system representation and address long-tail user needs. Finally, automating the detection of hallucinations and misinformation can create healthier information environments for society.

6.6 Broadening the IR Community

Expanding the scope of IR research relies on insights from multiple disciplines. Psychology can contribute to developing realistic and diverse user profiles, ensuring that simulations reflect a wide range of user behaviours and needs. HCI can enhance these efforts by validating simulated models against real-world user data, helping to identify and correct any shortcomings. In social science and policy research, simulated users can be used to audit digital systems, for example, by emulating typical search activities on platforms like podcasts or social media. This approach may help researchers better understand how algorithms shape user experiences. Additionally, techniques inspired by behavioural economics may highlight how people interact with online systems, providing valuable input for refining evaluation models in IR. By drawing on expertise across these fields, the IR community can build more comprehensive models of user behaviour and create more robust evaluation frameworks. This interdisciplinary approach ensures that IR systems are designed to meet the needs of diverse populations and adapt to changing technologies.

6.7 Obstacles and Risks

LLM-based simulation for information access evaluation faces several obstacles. For example, validation costs are significant as large-scale evaluations are necessary to validate simulations for each component, and these results may not generalise across different domains and cultures. Trust is another critical factor; for LLM-based simulations to be widely accepted as a standard evaluation methodology, rigorous experiments, follow-up studies, and extensive discussions are required. The accuracy of these simulations is also limited by the capabilities of LLMs and simulation methods, meaning that the success of LLM-based simulations is heavily dependent on the evolution of LLMs.

There are also risks associated with LLM-based simulations. Self-training biases may arise as LLMs are increasingly trained on LLM-generated data, potentially leading to concept drift and system degradation. Self-preference bias is another concern, where LLMs may favour LLM-generated content over human-generated content, skewing evaluation results. Additionally, LLMs trained on existing test collections may lead to test data leakage, inflating evaluation results. The reliance on LLM simulations could result in a loss of diversity in evaluations, as they may not account for creative or unconventional outputs. There is also a risk of systems being optimised to perform well against LLM-based metrics but poorly under manual judgements, leading to overfitting to LLM-simulated relevance data. Furthermore, the validity of empirical conclusions is limited to currently available LLMs, and content-injection attacks pose a threat to the reliability of LLM simulations. Decentralisation of simulation methods could lead to untrustworthy results, emphasising the need for community governance. Lastly, the deluge of data generated by this approach necessitates proper methodological guidelines to ensure meaningful evaluation results.

7 Centering Societal, Democratic, and Emancipatory Values and Ethics in IR

7.1 Description

The core of social responsibility, ethics, and environmental impact in IR and technology revolves around the values held by individuals, communities, and societies and how these values are prioritised. The IR research community must consider whether its values prioritise algorithmic improvements over the ethical and societal impacts of systems. Often, these values are implicit and not explicitly justified. This section explores the diversity and role of values in IR, proposing a research agenda to understand and justify these values and to develop methods for incorporating them into IR research and products. Values are considered broadly, encompassing human rights, ethical principles, social justice, environmental concerns, and practical desires like a search engine's ability to understand a user's language.

The explicit identification and critique of values are already practised in other computing fields and have a tradition in disciplines outside computing. Efforts to establish computing and information science values include Codes of Ethics like the ACM Code of Ethics and the International Federation of Library Associations and Institutions Code of Ethics for Information Professionals. The community is urged to *(i)* identify and make explicit the values governing IR work; *(ii)* engage in explicit debate and critique of these values; and *(iii)* develop techniques and research methods to understand and account for these values in developing, evaluating, and deploying IR systems. This approach should be technically, ethically, and sociologically rigorous, ensuring that IR systems serve their intended communities effectively and responsibly.

7.2 Motivation

Access to information is crucial to supporting informed citizenry in democratic societies and addressing global challenges like pandemics, conflicts, and climate change. The IR field is responsible for aligning research with societal needs, as emphasised by frameworks like "FACTS-IR" (fairness, accountability, confidentiality, transparency, and safety) and "IR for social good". However, focusing solely on technical aspects is insufficient; the IR community must acknowledge that research is influenced by political and social values, which should be made explicit and aligned with humanistic and emancipatory outcomes.

IR research already embodies values such as scientific rigour and access to information, but other values, such as aspirations towards artificial general intelligence (AGI) and monetisation, require scrutiny of their societal impact. Recent discussions suggest that democratic theory and value-driven perspectives can guide IR research towards social good, though these ideas remain peripheral. The IR research community is encouraged to critique and debate the values shaping IR, including foundational assumptions, to ensure alignment with desirable social outcomes.

7.3 Proposed Research

Following the approach used in outlining the research challenges in Section 7.4, we propose key research efforts to enable explicit engagement with values in IR across three levels: *(i)* community, which includes both the structures that support this research and studies of the research community itself; *(ii)* methodology, focused on developing techniques to identify and integrate values into IR practices; and *(iii)* application, which addresses the context-specific and domain-specific implementation of value-aware IR in various settings.

7.3.1 IR Community Research

To address the research challenges in the IR community, two main approaches are essential: reflective practice and external engagement with communities and stakeholders. Reflective practice involves creating opportunities for discussions on scientific, ethical, societal, and environmental issues among peers, including junior and senior members. This can be achieved through meetings, workshops, and panels, as well as by examining how values are characterised in current practices like Calls for Papers and reviewer instructions. Introducing reflective questions in the peer review process can also make these values more explicit.

External engagement focuses on centring societal, democratic, and emancipatory values by collaborating with other scientific disciplines and relevant communities, often overlooked in research processes. Insights from stakeholders on emancipatory and decolonising practices can help integrate these into IR research. Co-design and participatory approaches are crucial for aligning values across communities. For example, principles such as “Indigenous self-determination” and “Sustainability and accountability” from the Australian Institute of Aboriginal and Torres Strait Islander Code for Indigenous research emphasise the importance of Indigenous peoples’ full engagement in projects affecting them, ensuring their rights are respected per the United Nations Declaration on the Rights of Indigenous Peoples.

7.3.2 Methodology Research

Addressing the methodological research challenges in IR requires collaboration with scholars from various disciplines. The first step involves documenting values and processes within the IR community and those affected by it. This can be achieved through systematic reviews of IR literature to track value changes over time, examining documentation and marketing materials to understand value presentation, and conducting anthropological and ethnographic research to capture community perspectives. Additionally, HCI research can help understand user values and perceptions of

IR systems. Synthesising these values into coherent specifications while documenting tradeoffs and conflicts is crucial for methodological contributions that can be adapted across different contexts.

The design and research methods focus on involving non-experts in IR design and evaluation through participatory or co-design, allowing them to contribute their values directly. This approach, which includes *design futuring*, aims to reimagine new sociotechnical futures for information access. HCI offers methods to involve people in computing system design, which can be adapted for IR tasks. This participatory approach emphasises sharing power and decision-making with broader stakeholders, drawing on the rich knowledge from information science and HCI to centre human values in IR research.

Translation methods are essential for converting elicited values and design goals into technical implementations. This involves creating operationalisations of values with established technical implementations and developing methods to translate new values into usable models, algorithms, and metrics. Validating these implementations with the originating communities ensures they align with the intended values. Educational methods are also necessary to train community members in identifying and incorporating values into their IR work, making explicit the latent values in technical aspects like algorithms and evaluations. This educational effort spans all expertise levels, ensuring a comprehensive understanding of value integration in IR research and development.

7.3.3 Research on Specific Applications

Involving (Real) Users. Much existing IR research heavily relies on static datasets, which overlook the complexity and diversity of real users' behaviours, needs, and values. Future research should focus on designing IR systems tailored to distinct user groups with varying backgrounds. Conducting targeted user studies alongside static dataset experiments will provide nuanced insights into how different individuals and communities engage with these systems. This approach is crucial for investigating and measuring the impacts on real users. Additionally, it is important to provide mechanisms for end users to express their values and preferences without subjecting them to non-consensual live experiments, such as A/B tests. While direct user involvement in IR research can be challenging, leveraging LLMs to simulate users offers a promising alternative (see Section 6), though further investigation is needed to ensure these personas reflect diverse real-world user behaviours.

Understanding and Measuring the Impacts. IR research must develop robust methodologies to evaluate the long-term impacts of these systems at multiple levels—individual, community, and societal. This involves developing theoretical frameworks and empirical methods that assess immediate query responses and how system outputs evolve and influence user decision-making and social dynamics over time. The community must articulate stakeholders, such as platform owners, publishers, consumers, and society, and create spaces to critically examine, contest, and negotiate the tradeoffs between stakeholder needs. Given the ethical and logistical challenges of involving users in longitudinal studies, using LLM-based agents to simulate extended user interactions presents a promising pathway. These simulations can help identify phenomena like echo chambers or polarisation, yet further research is needed to scale these approaches and rigorously evaluate their validity and limitations.

Alignment with Values and Goals. Ensuring IR systems align with societal, ethical, and democratic values is crucial. Research should focus on creating simulation environments to test strategies for mitigating adverse outcomes like echo chambers and train models with reinforcement learning. Real-world monitoring tools should assess long-term social and ethical consequences, identifying and correcting misalignments with community goals. Developing intervention methods to fix misalignment problems is essential. Challenges in alignment are not only technological; research must consider how values are selected and negotiated, especially given power asymmetries between platform owners and users. Encouraging research on governance models for IR platforms is necessary to address these issues.

7.4 Research Challenges

IR Community Challenges. How research is conducted significantly affects the social and environmental impact of scientific contributions. Established practices often become norms that shape research operations, and there is a pressing need for open discussions about the values of the IR community and their societal and environmental implications. Key research questions include understanding the driving values of IR research, their historical evolution, and how to make these values explicit. It is crucial to consider the values held by users, artists, and others and how these may align or conflict with those within the IR field. Effective communication of these values to all stakeholders, including marginalised groups, is essential. Additionally, it is important to clarify the trade-offs made in research processes and to engage with Fairness, Accountability, Confidentiality, Transparency, and Safety (FACTS-IR) research. Measuring societal and environmental impact, conducting participatory research, and supporting social justice movements are vital considerations.

The impact of research findings is shaped by how research is defined, addressed, and conducted within the community, even when the focus is on theoretical or engineering challenges. Practical decisions, such as participant recruitment, often influence this impact. By explicitly acknowledging and reflecting on the values and the relationship between researchers and the potential impact of their work, a more open dialogue can be fostered, leading to a positive societal impact.

Methodological Challenges. Addressing the methodological research challenges in IR requires collaboration with scholars from various disciplines. The first step involves documenting values and processes within the IR community and those affected by it. This can be achieved through systematic reviews of IR literature to track value changes, analysing documentation and marketing materials to understand how values are presented, and conducting anthropological and ethnographic research to capture community perspectives beyond published artifacts. Additionally, HCI research can help understand user and creator values, perceptions of IR systems, and ideas for alignment. Synthesising these values into coherent specifications while documenting tradeoffs and conflicts is crucial. This work should focus on methodological contributions that can be adapted to different contexts, building a knowledge base for IR research that considers the values and needs of affected individuals.

Challenges on Specific Applications. The next challenge is designing and researching methods that allow non-experts to contribute to IR design and evaluation, bringing their values directly into the process. Participatory design or co-design of IR systems, including design futuring, can help reimagine new sociotechnical futures for information access. HCI offers methods to involve people at various participation levels, and adapting these to IR tasks will require identifying applicable methods and developing new ones. This approach should be grounded in specific applications and methodologically oriented to equip the IR community with skills for collaborative design. Participatory design reflects inclusivity and shared decision-making values, drawing on information science and HCI knowledge to centre human values in IR research.

Translating the elicited values and design goals into technical implementations is another significant task. This involves creating operationalisations of different values, allowing for established technical implementations to be adopted in development or research projects. Methods for translating new values into concrete implementations and validating them with community members are essential. For example, aligning democratic goals with diversity metrics in news recommendations demonstrates how values can be operationalised. This work produces reusable knowledge and artifacts, contributing to developing IR systems that reflect diverse values. Finally, educational methods are needed to train community members in identifying and accounting for values in IR work. This training should extend beyond those conducting user studies or participatory design to include making explicit the latent values in technical work on IR algorithms, models, and evaluations. By addressing these methodological research needs, the IR community can develop tools and systems that better align with the values and needs of the people they affect.

7.5 Broader Impact: Influence on Other Research Fields and Society

Bi-directionality in IR fosters interdisciplinary collaboration, integrating perspectives from fields like ML, computer vision, and NLP. This synergy is evident in adapting retrieval techniques in RAG for LLM-based models, which benefits IR and related fields. The convergence of these disciplines, particularly with applying LLMs across data types, enhances opportunities for scientific exchange and collaborative efforts to address broader social and technological impacts. Such collaborations can lead to developing more inclusive and less biased systems, emphasising the importance of incorporating insights from diverse experts, including those outside the tech field. As IR technologies increasingly influence society, the field must acknowledge and address their impacts, such as shaping public opinion, increasing polarisation, and spreading misinformation, disproportionately affecting marginalised groups.

7.6 Broadening the IR Community

The field of IR has increasingly focused on the technical challenges of building and evaluating systems such as search engines and recommendation systems over the past 40 years, leading to significant societal changes. However, this focus has created a disconnect between those who build systems and those who understand their broader impacts. Consequently, issues related to the social impacts of IR systems are often addressed in other fields, such as AI ethics, information science, and digital ethnography. To bridge this gap, IR should integrate methods and ethical frameworks from these fields, requiring a re-expansion of its methodological repertoire to include mixed methods and qualitative research. This shift will necessitate changes in reviewing practices and a greater appreciation for diverse research approaches.

Assessing the societal impacts of IR work is challenging, as current practices prioritise measurable and quantifiable metrics, often overlooking diversity, misinformation, environmental impacts, and minority exclusion. New metrics are needed to evaluate these aspects, but they should not be the sole indicators of value. IR must engage with these assessments by creating opportunities for participatory research and ensuring the inclusion of diverse voices, adhering to principles of self-determination and sustainability. The call for explicit discussion and negotiation of values aims to foster inclusive and rigorous practices, enabling the IR community to better understand and address shared and divergent values and methods.

7.7 Obstacles and Risks

Integrating explicit research values within the IR community presents several challenges. Technically, adhering to these values may limit researchers' choice of solutions, potentially leading to less effective systems and difficulties in publishing at prestigious venues. Engaging the IR community is another hurdle, as researchers may perceive their work as neutral and struggle to see the relevance of values, risking community fragmentation. This division could extend to industry, where differing values might hinder collaboration, such as sharing datasets and internship opportunities.

Externally, the IR community risks isolation if it fails to engage with explicit values, potentially stifling innovation and limiting its impact compared to other fields that embrace such values. Balancing core values without obstructing progress is crucial, as seen in some European countries where AI regulations have paused technological advancements. Evaluating the success of a values-based framework is essential to avoid it becoming a bureaucratic burden, with considerations including societal and technological impacts. Accountability is another concern, as defining universally acceptable values is challenging, and evolving societal impacts of technology may lead to discrimination or exclusion of certain research topics.

7.8 Positionality Statement

The team responsible for this section of the report is diverse in gender, cultural background, and disability, with expertise spanning computer science and information science. However, there is an acknowledgement of the lack of representation from other relevant communities and disciplines. The report was drafted on Wathaurong Country, unceded land in Australia, with a rich history of over 40,000 years. The report calls for creating spaces and removing barriers to include broader perspectives in IR discussions, including those from academia, user communities, and journalism. Workshops are highlighted as effective venues for such discussions, provided they are designed to be accessible. Policies like low-cost workshop-only registration can facilitate participation for those without computing research travel budgets. Additionally, conferences should aim to connect with relevant communities and activities in host locations to foster mutual benefits beyond tourism.

8 Evaluation of Complex IR

8.1 Description

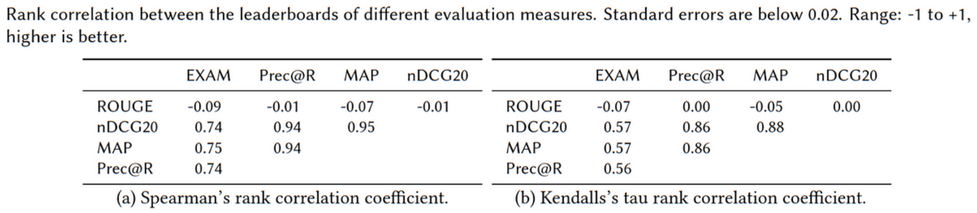
IR has evolved from document retrieval to dynamic information delivery, driven by advancements like LLMs. Unlike traditional search systems that provide repeatable document-based results, LLM outputs are variable and non-repetitive, challenging conventional evaluation methods. As IR systems increasingly integrate AI-generated content with traditional search results, future evaluation must focus on whether the right information—regardless of format—is provided rather than simply the right documents. Emerging models may include domain-specific LLMs or interactive platforms where users engage directly with topic-focused conversational agents. These developments call for new, reproducible evaluation methodologies that address modern information delivery's dynamic and multimodal nature.

8.2 Motivation

Evaluating the quality of System Output Packages, including a mix of generated and static multimedia content, presents unique challenges and opportunities. Traditional Cranfield-style evaluations are less effective in this context because the content is no longer static, and system responses include dynamic elements such as fragments, answers, generated paragraphs, images, videos, and graphs. These elements can be ephemeral, rendering judgments meaningless shortly after they are made. The concept of relevance in System Output Packages should be broader than the traditional Cranfield focus on topical relevance, incorporating new variables like the amount of information, specificity, and creativity. This shift suggests that relevance should be a multidimensional construct, moving away from the simplicity of a unidimensional scale of topical relevance.

Several principles and assumptions guide one possible approach to evaluating System Output Packages. Humans must be involved in the evaluation process, as information access systems are ultimately intended for human use. Each interaction with the system potentially returns a different response, indicating the absence of a static corpus for manual inspection. Despite this, evaluations should remain reproducible and reusable, necessitating a reevaluation of test resources since rerunning a system results in different System Output Packages. Additionally, evaluations should provide insights into which parts of a system response are relevant rather than offering a global rating for the entire response. This approach emphasises the importance of explainability and partial relevance in the evaluation process.

Where we are today. The dominant model of offline evaluation in IR, using test collections from campaigns such as TREC and NTCIR, relies on foundational assumptions that are increasingly outdated. Online evaluation, while an alternative, is impractical for academic research due to its requirement for substantial user traffic, limiting its feasibility to high-volume commercial search systems. Observational and interventional user studies offer another evaluation paradigm but are often small in scale and demand significant effort for each experiment. While still relevant, these methods do not adequately support the research community outside the industry. The traditional Cranfield approach to IR evaluation abstracts the search process with assumptions that simplify evaluation, yet are unrealistic. Ellen Voorhees of TREC identified several assumptions: users search static collections of documents, their information needs remain static during a search, and



Rank correlation between the leaderboards of different evaluation measures. Standard errors are below 0.02. Range: -1 to +1, higher is better.

	EXAM	Prec@R	MAP	nDCG20
ROUGE	-0.09	-0.01	-0.07	-0.01
nDCG20	0.74	0.94	0.95	
MAP	0.75	0.94		
Prec@R	0.74			

(a) Spearman's rank correlation coefficient.

	EXAM	Prec@R	MAP	nDCG20
ROUGE	-0.07	0.00	-0.05	0.00
nDCG20	0.57	0.86	0.88	
MAP	0.57	0.86		
Prec@R	0.56			

(b) Kendalls's tau rank correlation coefficient.

Figure 3. Leaderboards under a Question-based nuggets (EXAM) are highly correlated with human judgements, however the ROUGE metric is uncorrelated. Study on TREC CAR Y3 data. Figure from Sander, David P., and Laura Dietz. "EXAM: How to Evaluate Retrieve-and-Generate Systems for Users Who Do Not (Yet) Know What They Want." DESIRES. 2021.

relevance is defined topically. Additionally, effectiveness is measured by precision and recall, and all documents are assumed to be equally accessible and recognisable by users.

In response to evolving system responses, alternative evaluation methods from other fields, such as the summarisation community, are considered. Metrics like ROUGE and METEOR, which rely on fuzzy word matches and penalise changes in word order, are used to compare generated responses to gold-standard responses and are designed for short answers of well-defined information needs ("What is the capital of Australia?"). However, these metrics lose efficacy with longer responses due to differences in word choice and spurious matches (see Figure 3). A promising alternative is nugget-based evaluation, where the information need is broken into smaller components, each representing an important fact. The more nuggets covered, the higher the quality of the response. Although nugget-based evaluations have traditionally been manual and costly, recent research shows that nugget-based leaderboards correlate well with human judgements, offering a robust starting point for evaluating generated information objects.

8.3 Proposed Research

The study of evaluation metrics from a complementary perspective involves several key research directions. Developing integrated benchmarks is crucial for applying multiple evaluation metrics to the same collections, covering a broad taxonomy of relevance dimensions for comprehensive system evaluations. Statistical analysis of metrics can uncover important relationships, such as correlations and subsumption relationships, guiding efficient evaluation strategies. Thus, developing metric diversity-oriented significance tests ensures that system improvements are robust and not metric-specific. Correlating metrics with human assessors is essential to determine which aspects of relevance each metric captures and to identify gaps where metrics fail to reflect user-perceived quality. Additionally, improving source-based metrics is necessary to address challenges like hallucination and creativity in generated content, ensuring their applicability across various domains. These research directions aim to create comprehensive, reliable, and multidimensional evaluation frameworks for modern IR systems, with community collaboration being vital for developing shared resources that ensure evaluations are rigorous, reproducible, and aligned with real-world user needs.

8.3.1 Better Nugget-Style Grading Rubrics

Current research on grading rubrics and nuggets is centred around a *bag-of-relevant-pieces* approach, with challenges needing attention. One challenge is ensuring system responses provide utility beyond merely restating the information need, as IR often focuses on matching queries to responses. Another challenge is improving the coverage of nugget rubrics, either biased towards a human judge's world knowledge or inspired by system responses in a judgement pool. Exploring methods that ensure nugget rubrics are comprehensive and serve diverse user groups effectively is crucial. Additionally, handling matches of different granularity and collating fine-grained nuggets into broader themes is essential. As coverage improves, multiple themes may emerge, each subdivided into detailed nuggets, requiring consideration of how short responses might highlight only some themes while longer responses could cover more detailed nuggets. Lastly, measuring the outline structure of system responses, particularly for topical coherence and the order of nuggets, is vital. As responses grow longer, maintaining topical coherence and observing the correct order in processes or historical events becomes increasingly important.

8.3.2 Study How People Search in Generated Multimodal Information

A large-scale, community-driven data collection effort is essential for understanding how users interact with new hybrid and dynamic IR systems. Traditional evaluation methods, which focus on the relevance of static documents to queries, are inadequate for future systems where responses are generated and vary by user, context, and time. To address these questions, researchers need datasets with queries, context, and interaction to identify features that influence user satisfaction, engagement, and trust. Such research seeks to uncover what motivates new IR systems, effective presentation styles, reasons for user abandonment, and engaging modalities. The effort requires diverse data collection across regions, economic classes, age groups, and cultures, necessitating significant funding and leadership from global industrial, governmental, and academic research labs. Success in this endeavour could lead to a deeper understanding of user goals, response types, demographic interactions, and preferences in IR systems while highlighting the importance of intersectionality by revealing how overlapping social identities and backgrounds shape user experiences and needs.

8.3.3 The Formal Nature of Information vs. Document

We propose developing evaluation frameworks that differentiate between content relevance and the appropriateness of the amount of information returned. Current methods often overlook how much information is optimal for a given user and query. To address this, evaluation benchmarks should model users' information needs and prior knowledge, recognising that certain information may be redundant or genuinely informative depending on the user. Additionally, we aim to create benchmarks that assess core competencies like reasoning and inference within information access contexts, an area largely neglected, as most current datasets focus on multiple-choice or logic-based tasks. Another crucial dimension is information expectedness: incorrect but expected information can be especially harmful, so evaluations should consider how anticipated content influences user trust and comprehension. This aspect, tied to issues like misleadingness and hallucination, can be effectively explored through targeted user experiments.

8.3.4 Adaptive and Context-aware Evaluation in Proposed Research

As IR evolves beyond static document lists towards dynamic aggregations and context-aware presentations, there is a need for an adaptive and context-sensitive evaluation framework. Traditional IR metrics (e.g., NDCG@k, Average Precision, RBP, Precision@k) are grounded in user models that assume users browse through documents presented in order of relevance or within a defined top-k set. These approaches enable controlled experimentation but rely on the premise that users encounter well-contained items independently. However, modern systems increasingly present information synthesising content from multiple sources and adapting to individual user contexts. This shift introduces greater variability in content aggregation and presentation methods, complicating the evaluation process by adding factors that traditional metrics cannot adequately address. Future research should focus on establishing robust and repeatable offline evaluation methods suited to these emerging systems. Rather than measuring only topical relevance against an information need, new frameworks must account for multi-dimensional aspects of user satisfaction relative to their specific tasks. Key priorities include defining which attributes, such as relevance, alignment with user intent, creativity, and degree of information gain, should be measured and determining how best to quantify them across diverse formats. The goal is to develop methodologies agnostic to the delivery mode while representing different users' abilities, tasks, and comprehension levels. Research can create meaningful tools for optimising future IR systems in increasingly complex environments by articulating these measurable attributes and ensuring complementarity between evaluation metrics.

8.3.5 Whitebox Evaluations for Richer Relevance Labelling

We propose developing relevance labelling methods that capture richer, more nuanced relevance aspects, such as the location of relevant information within a response, the reasons for its relevance, and the specific issues covered. Current test collections primarily use overall relevance judgements (e.g., relevant, not relevant, or graded relevance), which offer limited insight into which parts of a response are valuable and why. To address this, whitebox relevance labelling can provide a deeper understanding of system error modes and improve ML-based optimisation beyond blackbox reinforcement learning methods. Prior approaches, like marking non-relevant segments or highlighting key information, have been hindered by inconsistent judgements across assessors. We propose leveraging AI and LLMs to standardise relevance annotations, with human judges verifying AI-generated rationales to ensure quality and consistency. A significant challenge is the potential increase in annotation costs. We suggest non-intrusive data collection methods similar to interaction data to mitigate this. For example, judges could indicate the first point in a response where they determined relevance, allowing an LLM to infer and propose a rationale. This process would streamline the creation of structured relevance rubrics while ensuring minimal overhead for assessors. This research aims to create more informative, cost-effective relevance annotations that better support the evaluation and improvement of modern information access systems.

8.3.6 Lampedia: A Corpus of LLMs

It is plausible that the "information" may take additional new forms. Currently, LLMs are primarily tools for engagement with the collected information on the Web. They are notoriously unreliable, not only confabulating but also bringing material from unrelated topics to their answers. Research and experience show that LLMs can be more trustworthy when purpose-built for specific domains. Thus, it is plausible that information providers will begin by identifying an appropriate LLM for a topic and then either return an answer from that LLM or even possibly return a link to the LLM to allow the user to explore their topic conversationally.

Thus, a possible future resource is the emergence of a Wikipedia-like collective of LLMs, “LMpedia” say, in which every “page” is an LLM on a specific topic, each one maintained by editors in how Wikipedia pages are managed today. The task of “search” would include identifying the right LMpedia page - a process that helps ensure topical focus and accuracy and maintains a key information resource in a public form. The subject of this section is how to undertake reproducible evaluation in an environment of dynamic information. New evaluation methodologies will need to embrace such developments.

8.4 Research Challenges

8.4.1 How Evaluation Changes from the TREC Model in the New World

Evaluating generated content presents unique challenges. Simple rating methods, while straightforward, lack diagnostic detail, offering little insight into where specific issues occur within the output. Identifying precise flaws is especially difficult compared to evaluating static documents in traditional IR systems. A key challenge lies in redefining relevance beyond a single dimension. While past frameworks focused on topical similarity or functional inclusion, modern evaluation must embrace a multidimensional approach. Relevant dimensions include trustworthiness, completeness, comprehensibility, and coherence—each potentially applying to entire outputs or individual components. Balancing these dimensions without overwhelming assessors is a significant hurdle. Another challenge is reconsidering the independence assumption in relevance judgements. Traditionally, documents were evaluated in isolation for efficiency, but dynamic content may require comparative methods, such as pairwise assessments, to capture nuanced differences in relevance. Exploring how to integrate these complex evaluations into scalable frameworks remains an open research question.

8.4.2 Evaluation Metrics as Partial Evidence

The evaluation metrics proposed for dynamic, generated content are highly diverse, capturing various aspects of relevance such as overlap with a gold standard (e.g., ROUGE, BLEU), consistency with sources, readability, and interpretability. This diversity highlights the need for evaluation benchmarks encompassing these complementary relevance dimensions, ensuring system optimisation does not overly focus on a single aspect. Complementarity among metrics arises from the aspects they measure and their differing strengths and weaknesses as predictors of relevance. Metrics based on information unit overlap are objective and easily interpretable, but struggle to capture abstract similarities between outputs and gold standards. Conversely, metrics leveraging processing tools or language models better grasp content abstraction yet risk introducing significant bias. The core challenge lies in developing evaluation benchmarks that effectively balance these trade-offs, leveraging the strengths of various metrics while covering the full spectrum of relevance considerations.

8.4.3 Evaluation from Social Perspectives

IR system evaluation raises as many social concerns as their development. While traditional evaluation focuses on static documents, modern systems deliver diverse, context-sensitive, and personalised outputs, making universal performance metrics like relevance judgements less effective. To address this complexity, user involvement in the evaluation process is essential. Cultural diversity must be a key consideration, ensuring consistent information quality across user groups. Since LLMs may reflect cultural biases in their training data, evaluation frameworks should identify and mitigate risks of generating offensive or misleading content. Evaluation transparency is equally important. Users interacting with systems may receive varying outputs; therefore, mechanisms should explain how and why information differs, fostering trust and understanding. User collaboration in developing evaluation metrics is crucial, especially for domain-specific systems where expert input enhances relevance and fairness. Building pipelines for user contributions can help shape evaluation frameworks that better reflect diverse needs. Though some of these concerns have been explored—such as fairness in argument retrieval—dynamically generated content requires revisiting past assumptions. Social perspectives in IR evaluation present rich opportunities for future research and development.

8.4.4 Information vs. Document

The shift from static documents to dynamic, system-generated content transforms the evaluation process, introducing new considerations for previously stable features. Evaluating relevance now involves identifying the correct document and determining the specificity and quantity of information, as relevance is influenced by the user's existing knowledge, making overly familiar information less valuable. Metrics must assess the alignment of system outputs with their information sources, focusing on truthfulness, hallucination detection, and creativity, though evaluating source-based accuracy remains challenging. Dynamic systems often provide inferred or implied information through reasoning, complex database queries, and structured data processing, making evaluating a system's reasoning and interpretation capabilities essential yet challenging. Additionally, attributes such as comprehensibility, coherence, and interpretability have become critical regardless of the query, unlike in traditional document retrieval, where these factors were often overlooked. Beyond accuracy, evaluators must also consider the plausibility of generated false information, as the risk lies in generating errors and producing content that appears deceptively credible.

8.4.5 Nouvo IR

Nouvo Information Retrieval (NIR) systems surpass traditional IR by crafting tailored responses to user requests through context-aware interactions, aiming to create an "ideal relevant document" rather than retrieving existing ones. This evolution presents complex evaluation challenges, such as determining the composition type (concise answers, curated summaries, SERP-style pages) and the evaluation unit (citations, generated documents, information nuggets). Evaluating these outputs requires attention to personalisation, accuracy, coverage, and clarity. For instance, a factual request like "What is the deadline to drop a course?" demands accuracy and citation, while a complex inquiry like "Compare the Master's in Computer Science and Data Science programs" requires a detailed comparative table evaluated for completeness. Innovative requests, such as "Create a timeline of university research breakthroughs", necessitate assessing information accuracy and visualisation effectiveness. These examples highlight the need for flexible, multidimensional evaluation frameworks to address NIR systems' diverse response types and evolving user needs.

8.4.6 Challenges in Evaluating an Agentic Information Retrieval (AIR) System

Evaluating advanced information retrieval (AIR) systems presents innovative challenges beyond traditional methods. Key issues include assessing dynamic relevance for generated content without fixed test collections, ensuring faithfulness to source data while detecting hallucinations, and maintaining response consistency over time, especially with personalisation. Evaluations must consider multi-faceted responses that balance comprehensiveness and conciseness, as well as interpretability through transparent system explanations. Adaptability to user preferences, efficient response generation, and fairness in source representation are crucial. Scenario-specific challenges include handling open-ended, comparative, and multi-intent queries, as well as ensuring personalised and credible responses.

8.5 Broader Impact: Influence on Other Research Fields and Society

As IR systems increasingly depend on generated content, ensuring trustworthy evaluation is crucial not only for IR research but also for fields like education and psychology, where the impact of information access on cognition and learning is studied. Generated outputs pose challenges in reliability and trustworthiness due to a lack of transparent sourcing, necessitating new validation standards grounded in multidimensional evaluation methodologies. Reproducibility is a significant concern, especially when evaluations rely on LLMs that undergo retraining, risking scientific integrity due to potential variability in results. While automated metrics are useful, human judges are essential, as traditional frameworks like the Cranfield paradigm and user-based methods such as A/B testing are inadequate for large-scale evaluations. Developing new, efficient, and human-centred evaluation solutions is urgent to prevent IR systems from failing to assist users effectively, thus undermining the purpose of information access technologies.

8.6 Broadening the IR Community

The challenges in evaluating modern IR systems extend beyond the IR field, posing significant implications for the broader research community. Disciplines such as education, psychology, and behavioural sciences increasingly rely on accurate information access to study cognition, learning, and decision-making. If evaluation methods fail to ensure IR systems' reliability, reproducibility, and trustworthiness, research in these fields risks being built on unstable foundations. Inconsistent or biased information delivery can compromise experimental validity, skew findings, and hinder the development of evidence-based policies. Moreover, the inability to reproduce evaluation results, especially when reliant on evolving LLMs, threatens the core scientific principle of replicability, potentially eroding trust across interconnected research domains.

8.7 Obstacles and Risks

Pursuing innovative research in evaluating modern IR systems faces significant challenges, primarily due to the high costs of running LLMs and the financial and logistical burdens of involving human judges. Privacy concerns and legal restrictions limit access to user data, hindering realistic evaluations. Additionally, the pressure for rapid method development may lead to shortcuts that compromise thorough evaluation and long-term reproducibility. These challenges threaten the development of trustworthy and user-centric IR systems. To address these issues, the research community is urged to collaborate on developing shared evaluation resources, open benchmarks, scalable evaluation frameworks, and community-driven standards to ensure robust, reproducible, and human-focused IR evaluation methods.

9 Minor Topics

9.1 Adversarial attacks to Information Access Systems

Information access systems have long been subjected to adversarial attacks, e.g., adversarial search engine optimisation. The uptake of generative LLMs for powering information access systems exposes these systems to new attacks that bring new challenges. These include misinformation injection, coordinated multimodal attacks involving text and images, data poisoning in training and retrieval corpora, and private data extraction through prompt engineering. LLMs also enable scalable, automated attacks, for example, generating large volumes of credible misinformation or simulating malicious user behaviours to compromise systems that learn from implicit feedback. Such attacks can undermine core information access infrastructure and impact downstream applications like RAG and retrieval-enhanced ML, with serious societal consequences, such as disrupting public health surveillance based on search analytics or influencing elections via manipulated results. The integrity of LLM-based evaluation methods is also at risk if synthetic data or simulated users are compromised. Addressing these threats requires research into robust defences, such as improved data provenance, anomaly detection, resistance to data poisoning, and privacy-preserving models. We encourage the community to investigate attack methods ethically and develop practical defences to ensure robust information access systems.

9.2 Cognitive Biases in the Era of GenAI based IR Systems

Cognitive biases affect how users interpret, choose, and trust information from IR systems. Biases like confirmation bias, anchoring, and the availability heuristic cause people to favour familiar views, focus on initial results or rely too much on easily found data. This narrows perspective and reinforces existing beliefs. To counter this, IR systems should show diverse viewpoints, be transparent about sources, and prompt users to think critically. LLM-based IR systems work differently from traditional ones. In essence, traditional IR systems match keywords and return links with little context. In contrast, LLM-based systems use deep learning to understand what the user wants, consider context and meaning, and pull together information from multiple sources for more tailored answers that adapt as the user interacts.

LLM-based IR systems can intensify cognitive biases by tailoring responses to reinforce users' beliefs and habits. For instance, when these systems prioritise information aligned with a user's past views, they strengthen confirmation bias and limit exposure to different perspectives. Users may also become anchored to the first answer provided and fail to consider alternatives critically. LLMs risk amplifying the availability heuristic by favouring trending or easily found content over less visible but authoritative sources. How information is framed can further influence interpretation and confidence, while overemphasising well-known or recent sources encourages authority and recency biases. If the underlying data reflects dominant opinions or stereotypes, groupthink can be reinforced, and harmful generalisations perpetuated. Together, these effects can narrow understanding, reduce exposure to diverse viewpoints, and potentially mislead or manipulate users without them realising it. These biases can interact in different ways. Users may already have tunnel vision due to their biases, while the system can reinforce or exploit these tendencies to influence or manipulate the user.

Tunnel Vision Example. Tunnel vision in LLM-based IR systems happens when users focus only on information that fits their existing views. For example, a user who believes climate change is purely natural may keep asking for responses supporting this idea. As the system learns from these repeated queries, it may prioritise similar sources and arguments, creating a feedback loop where only one side of the debate appears. This confirmation bias limits exposure to other perspectives. In addition, availability bias can worsen if the system pushes frequently accessed content over less common but important viewpoints. Users who demand answers from specific ideological or regional sources, like only conservative news outlets, further restrict the diversity of information they see. This narrow focus not only reduces understanding but also strengthens groupthink and stereotypes.

Agent/Search Engine Manipulation Effects (AME/SEME) Example. LLM-powered IR systems can unintentionally exploit cognitive biases and steer users toward certain viewpoints, a phenomenon known as Agent/Search Engine Manipulation Effects (AME/SEME). For example, anchoring bias makes users trust the first answer they see, especially if it matches their past behaviour. If someone often clicks on sources supporting stricter immigration policies, the system will likely keep showing similar content. This reinforces confirmation bias and narrows the user's understanding over time, as they mostly see information that fits their beliefs. Recency bias can make things worse by pushing newer or more emotional stories to the top, further shaping opinions based on what is most attention-grabbing rather than what is balanced or comprehensive. Together, these effects can distort how users see complex issues by exposing them mainly to one-sided perspectives. To address these challenges, LLM-based IR systems should present diverse viewpoints, clearly show sources, and prompt users to think critically.

Key Questions.

- How can LLM-based IR systems be designed to reduce confirmation and anchoring biases, ensuring users are exposed to diverse perspectives?
- What methods can be employed to balance the visibility of popular versus authoritative sources, avoiding the dominance of easily accessible or trending information?
- How can the framing effect be mitigated in LLM-generated responses to prevent users from misinterpreting nuanced or uncertain information?

- What strategies can be used to address authority bias, ensuring that emerging or unconventional viewpoints are given adequate attention?
- How can LLMs be trained to minimise recency and groupthink biases, promoting a more balanced representation of past and present knowledge?
- What are effective approaches for detecting and mitigating stereotyping biases in LLM-based IR systems, particularly when handling sensitive topics or data?
- How can user interaction and feedback mechanisms be integrated to continuously improve the system's ability to counteract cognitive biases over time?

9.3 Conversational Information Access: Multi-Agent Reasoning, Multimodal Understanding, and Adaptive Evaluation

Generative conversational information systems are transforming how people access information, moving beyond simple text interactions to support multimodal interfaces and new tasks. Since SWIRL 2018, researchers have developed early systems that added conversational features like query rewriting and passage retrieval to traditional IR models. These advances led to commercial adoption in 2022 with products such as ChatGPT and other generative search tools. Alongside this progress, the community introduced benchmarks like TREC CAsT, TREC iKAT, and QReCC, creating new ways to evaluate these systems. Despite these steps forward, progress has been gradual. The ambitious vision outlined at SWIRL 2018 is still not fully realised. To move beyond incremental gains and achieve a leap in conversational assistants, the field needs a new wave of LLM-based architectures designed for richer conversation and broader capabilities. This will require more realistic benchmarks that capture real-world complexity, dynamic evaluation methods that reflect ongoing interaction, better use of conversational context, and support for advanced multimodal features.

Key Challenges.
Architecture:

- Building conversational agents, encompassing features and components identified in the literature, such as failure mode, engagement, and Generated Information Objects (GIOs)
- Developing the next-generation architecture to perform end-to-end optimisation of all modular components in a conversational system

Evaluation:

- Developing reliable and informative methods for evaluating dynamic trajectories of conversations at scale, including both human and automatic evaluation.
- Proposing resources and formal methods for meta-evaluation. As novel approaches to evaluating conversational information access emerge, new benchmarks and formal methods are required to evaluate automatic evaluation techniques and systematically characterise the dimensions they support.

User Modeling:

- Building user simulators informed by user models and evaluating them

- Utilising user models and digital twins in the optimisation process of conversational systems to improve personalisation and task completion efficiency

Personalisation, memory, and multimodality:

- New memory architectures for short- and long-term states to perform in-depth personalisation over a long user interaction history?
- Integrating multimodality into conversations to assist users in their daily tasks and life in both proactive and natural form

Conversational interaction is a fundamental property of information access systems that cross-cuts all of the ideas discussed in this report, such as evaluation with user simulation and next-generation architectures. As AI evolves towards more agent-like behaviour, these systems will become even more central and change how people interact with technology in new ways. We briefly outlined the challenges here, not because they are unimportant but because conversational capabilities are now assumed to be standard in GenAI systems. This report covers key issues such as architecture and evaluation with user simulation.

9.4 Discovering the Mechanisms of IR

Modern information access systems rely on complex models trained on vast user interaction data. As we move away from rule-based or heuristic systems toward these data-driven approaches, the inner workings of models become harder to understand intuitively. Mechanistic interpretation (i.e., understanding the internal workings and computations of neural models) aims to uncover which specific operations a model performs and how different components, such as those within a generative language model, work together during these processes. Several prominent techniques include: activation probing investigates which parts of a model activate in response to specific inputs; path patching explores how information flows through different pathways in the network; circuit analysis maps out how groups of components work together for particular tasks; and sparse autoencoders search for simpler structures hidden within complex models. These methods help build a clearer, human-understandable picture of how advanced AI systems reason and make decisions, highlighting their strengths, limitations, and potential biases.

An understanding of model internals helps identify and fix biases or bugs, as well as make targeted changes through model editing. Rather than focusing only on deploying general-purpose language models like transformers or reporting benchmark results, researchers should also investigate how these models work. Examining the specific algorithms will reveal why their performance improves or falls short, ultimately leading to better and more reliable systems.

Key Questions. Mechanistic interpretation is advancing rapidly, but most current research focuses on classification rather than ranking tasks; open questions remain for ranking. First, if traditional ranking signals are still important, how do they appear within complex neural models? Second, are these models discovering new signals that earlier methods missed, and how are those signals being used? Third, since smaller models can sometimes perform well at ranking tasks, can we identify and remove parts of larger networks that are unnecessary for ranking?

9.5 Freedom of Information Act Search

Analysing and retrieving government documents released under Freedom of Information Act (FOIA) requests represents a significant research topic. This area investigates methods for enabling transparent access to public records, addressing challenges posed by the large volume, complexity, and fragmented nature of FOIA-released materials. Research directions include developing techniques for collecting and preprocessing these records, such as segmentation, optical character recognition (OCR), and metadata extraction, before sensitivity review. Other key topics involve advancing search technologies capable of supporting domain-specific queries at various document levels (file, page, paragraph), facilitating multi-document linking and reasoning, and enhancing user experience through NLP tools for summarisation, text simplification, and conversational exploration. Additionally, ensuring provenance and attribution by connecting search results to sources is essential for supporting robust evidence-based reporting.

Research in this area is important for democratic accountability, such as enhancing journalism through access to government records. In addition, developing an openly available corpus furthers opportunities for advancing IR and NLP methodologies, offering a resource for experimentation and benchmarking. This research area could foster community-driven collaboration by bringing together government agencies, NGOs, journalists, and researchers to promote transparency via open-access initiatives. Addressing scalability and sustainability is another important facet, as shared research infrastructures are needed to manage the continuously expanding government data. However, several obstacles persist: managing and indexing these vast datasets presents ongoing technical challenges; existing search tools often fall short in supporting complex multi-document evidence synthesis; and there is a growing need for IR models that move beyond simple keyword matching or generic AI responses to deliver precise, source-grounded answers. Overcoming these challenges remains essential for the full impact of transparent access to public records in support of informed civic engagement.

Key Challenges.

- **Massive and complex data:** Government data includes emails, memos, and official decisions, often redacted and dispersed across millions of documents.
- **High recall and precision demands:** Journalists require exact information for investigative reports, necessitating sophisticated IR tools.
- **Evolving information needs:** Queries are dynamic, requiring iterative searches and synthesis of evidence from scattered documents.
- **Existing solutions are inadequate:** Traditional search engines and chatbots lack the depth and precision needed for FOIA investigations.

FOIA Search is a pioneering effort to merge cutting-edge IR technology with societal needs. This initiative aims to ensure journalists and citizens can effectively access and analyse government information by overcoming technical and usability challenges. This research topic will set new standards for transparency, accountability, and public engagement in open government data.

9.6 Future of IR: Mars Shot to Information Access

When we think about the future, we think of flying cars, shining glass cities, perfect clean air, and advanced technology that integrates seamlessly into our day-to-day lives. We may think of systems as seen on science fiction shows such as Star Trek, where intelligent computers respond instantly to human queries and where access to vast repositories is effortless. Information access technologies, systems that retrieve data and interpret, contextualise, and personalise it to meet individual needs in real-time, will be central to our future, be it a utopian or a dystopian future.

Utopia: Towards Perfect Knowledge Democracy. Future information systems must be open, inclusive, and decentralised to support democratic access to knowledge. Open access platforms, academic repositories or community-run databases are key to breaking down barriers and reducing educational and economic inequality. However, access alone is not enough. Interfaces must adapt to different users. Modular and personalised designs should accommodate diverse learning styles, languages, and accessibility needs. Adaptive technologies and user-centred design can make systems more effective for all. Current knowledge systems are dominated by a few centralised platforms, such as major search engines and social media companies. This centralisation concentrates power, introduces algorithmic bias, and limits access to diverse perspectives. To counter this, we need decentralised models, supported by secure and verifiable data storage, to ensure transparency, traceability, and shared control over information. We must also rethink what counts as information. Increasingly, knowledge comes in dynamic, personalised, and short-lived data, not just static texts or documents. This shift challenges how we store, retrieve, and validate information and requires new frameworks and technologies that can handle this complexity.

Key Questions. What is stored and how? What is information in the future? Who decides what knowledge is valuable or authoritative? What role should users play in shaping their information environments? How do we ensure transparency without compromising individual privacy?

Dystopia: The Fractured Information Future. At odds with the information utopia is the dystopia, marked by manipulation, misinformation, inequality, and a lack of transparency, privacy, or accountability. **What could go wrong? Everything.** Surveillance is everywhere in this dark future. Users are tracked, while their data is harvested and sold without consent. Algorithms, optimised for profit and control, manipulate public opinion and obscure reality. Truth becomes elusive, buried under misinformation, deepfakes, and synthetic content designed to deceive. Access to information becomes a privilege, not a right. Wealthy individuals and their organisations hoard data, while underprivileged communities are excluded. Personalised interfaces become echo chambers, reinforcing biases and suppressing critical thinking. Centralised systems consolidate power in the hands of a few, unchecked and unaccountable. Without transparency, algorithms act as unexplainable, inscrutable, and unchallengeable black boxes. Decentralisation remains an abandoned ideal, with users left powerless. Instead of empowering users, information systems become tools of control, surveillance, and inequality, a fractured, volatile ecosystem with devastating consequences.

Key Questions. Who controls the data and algorithms? How can users verify the truth in a sea of misinformation? What protections exist against surveillance and exploitation? How do we prevent the erosion of public trust in knowledge? What happens when knowledge access is driven solely by profit and power?

9.7 IR With Low Resource Languages

IR research traditionally focuses on languages that are widely spoken, have well-documented linguistic resources, and hold strong societal representation. In contrast, low-resource languages, which have limited data, processing tools, or institutional support, are often overlooked in IR research. This imbalance leads to barriers to information access for these languages.

Conducting IR research with low-resource languages presents numerous challenges. These include limited digital and linguistic resources, the necessity of collaborating with subject-matter experts or protected groups such as Indigenous communities in remote areas, or a general lack of computational resources and digital infrastructure for users. Together, these barriers create a cycle where limited resources hinder research progress, reducing the visibility and usability of these languages within information systems. Despite these obstacles, IR has the potential to break this cycle by improving access, visibility and inclusion for low-resource languages in the digital sphere. Initiatives such as CLEF, NTCIR, and FIRE have demonstrated that the IR community can broaden its scope beyond TREC's historical focus on English. These efforts have expanded research to include European, Asian, Indian, and African languages, but most of the world's spoken languages remain unrepresented in IR research. This is often due to limited involvement from speaker communities in shaping IR systems and benchmarks. Rather than unintentionally contributing to language erosion, IR research should actively support the safeguarding and revitalising of low-resource languages. This can be achieved by pursuing new initiatives, such as the Low Resource Environments Track at ACM SIGIR 2025, and embracing participatory research alongside co-design practices with language groups who benefit from our expertise.

Key Challenges.
Common IR challenges:

- How can we work with speakers of low-resource languages who are not yet part of the IR community? What are the participatory and co-design practices that we can adopt while adhering to the principles of self-determination, mutual benefit, and ethical research values?
- IR architectures that address data sparsity and evaluation methodologies tailored to low-resource settings.
- Do current IR evaluation metrics and test collections work well for these environments and languages? Are they biased toward well-resourced languages?

Language-specific challenges:

- Incorporating linguistic knowledge, such as oral information, grammar structures, and phonetic and morphologic variations.
- Respecting cultural and ethical protocols set by communities that own and speak these languages.

By understanding these challenges and opportunities, the IR community can play a key role in keeping low-resource languages alive and ensuring they are accessible and represented in digital spaces. The first step is to build a strong network that values the diversity of low-resource languages worldwide. This community could enable researchers and professionals to share their experiences, discuss what has worked (and what has not), and learn from each other when dealing with IR challenges in these environments.

9.8 Multimodal IR

The IR community has been engaged with multimedia and multimodal search research for decades. So why does this remain an important consideration for the coming years?

Combining different modalities leads to concepts that cannot be captured by text, image, or audio alone. Multimodal systems are fundamentally more complex than unimodal ones because they must integrate and reason across diverse information forms. Recent progress in encoder-decoder models and diffusion transformers has expanded what is possible with multimodal AI. These new systems support advanced conversational interfaces, retrieval tasks, and content generation applications that were not feasible just a few years ago. Looking ahead, users will increasingly express their information needs through multiple input types, not only text or speech but also images, sketches, video clips, music samples, touch, and gestures. Likewise, the data being indexed will span these same modalities: photos, diagrams, videos, audio recordings from meetings or podcasts, or musical pieces. IR systems must do more than return ranked lists; they must also generate rich multimodal responses that may include combinations of text, images, sound clips, or interactive elements. Users will anticipate seamless multi-turn conversations involving all available modalities. Addressing these challenges is essential for building future-ready information access systems. To truly support multimodal search and interaction, IR systems need to handle all aspects of working with different data types, not just text. This means they must be able to:

- Understand inputs in various forms such as text, images, audio, or video, including recognising different meanings or intentions behind each type and segmenting them properly to determine what the user wants.
- Index and encode information from multiple sources so that it can be searched using a unified framework.
- Retrieve and rank results that may come from different modalities or combine several media types in one answer.
- Generate responses using the most appropriate format for the user's needs, whether that means replying with an image, a piece of text, audio, or a mix.
- Adapt responses so they are easy to use on whatever device or platform the person is using, taking into account things like screen size or accessibility needs.

Many existing concerns outlined elsewhere in this report, especially efficiency (Section 2), architectures (Section 3), and evaluation (Sections 6 and 8), all need to account for this expanding set of multimodal user experiences. Multimodal experiences also support a more diverse and inclusive set of people, if thoughtfully designed. Processing costs are one risk, especially considering further pretraining or fine-tuning of these models. It is an exciting time where the field is developing and expanding rapidly with foundational abilities that would have been only dreamed of at previous SWIRL forums, and we should actively explore the opportunities that have opened up.

9.9 Neurophysiological IR

Wearable devices with built-in sensors are becoming increasingly common. Soon, earbuds may include electroencephalogram (EEG) sensors that measure brain activity. These developments offer both new opportunities and challenges for the IR community. Neurophysiological IR aims to use data from neural sensors (such as EEG or functional near-infrared spectroscopy) and physiological sensors (like electrodermal activity or galvanic skin response) to tackle problems like limited implicit user feedback and to enable new forms of brain-computer interaction in IR.

As information access moves beyond traditional search to include new modes like augmented reality and LLM-based conversational systems, real-time signals from wearable devices can play a key role. Devices like smartwatches or neural earbuds can collect valuable feedback, like emotional reactions or physical responses, while users interact with information access systems. For example, when someone uses a voice-enabled search assistant while wearing these devices, their implicit feedback can help the system better understand and adapt to their needs. This richer context allows for more personalised and responsive IR. However, physiological signals are challenging to work with. They generate large volumes of noisy data without clear meaning, making them hard to interpret and connect to user intent. Extracting useful information is also technically demanding, and collecting this data often needs specialised equipment and controlled environments. Since physiological data is highly personal, it raises privacy and ethical concerns.

Key Challenges. As neurophysiological IR continues to grow as an area, developing resources (datasets, benchmarks), methodologies, and/or best practices is a primary research challenge. Other research topics include:

- Enabling novel interactions and interfaces for information access, such as “querying/prompting by thinking” through brain-computer interfaces instead of typing or speaking, leading to novel applications and interactive scenarios.
- Measuring the user experience and understanding the user context during human-information interaction with rich neurophysiological signals.
- Evaluating information access systems using neurophysiological feedback.
- Replicability and reproducibility of experiments using neurophysiological signals.

Advancing these research topics will require close collaboration with other disciplines, such as neuro and cognitive science, psychology, and ubiquitous computing, among others.

9.10 Pervasive and Ubiquitous IR

While pervasive and neurophysiological IR (see Section 9.9) are related, they focus on different aspects. Both aim to make information access more adaptive and personalised by using signals beyond traditional queries, yet they differ in scope and emphasis. Pervasive IR uses a broad set of external signals from the user’s environment and activities to make information access ever-present and proactive. Neurophysiological IR relies on biological data from users as they interact with information systems. While both approaches can complement each other, since physiological data may become one source among many for pervasive IR, their focus and methods are distinct.

Pervasive IR is about embedding information access into everyday environments, devices, and activities. Its goal is to anticipate user needs and deliver relevant content without explicit search queries, so-called zero-query retrieval. Pervasive IR draws on contextual signals: device usage patterns, physical location, calendar events, environmental data from smart sensors, social interactions, or recent activity across platforms. The emphasis is on integrating situational context data to provide proactive assistance wherever the user is. However, collecting and sharing personal information across multiple devices raises privacy risks, including unauthorised access or unintended data exposure. When user data must be exchanged between agents or services to fulfil a request, there is an added risk of information leakage. One way to mitigate these issues is to use personalised on-device agents that process data locally and communicate externally only when necessary, following strict privacy protocols.

Key Challenges. One major challenge in this area is obtaining suitable datasets. Existing datasets from related fields, such as CLEF lifelogging, could be adapted and extended for pervasive IR tasks. Another approach is to develop reference systems with limited modalities or channels to enable focused data collection, particularly for studying information need prediction. Evaluation presents additional difficulties. It is important to consider what happens when a system predicts incorrect or incomplete information needs. Such errors can negatively affect user experience, undermine trust in the system, and may even be perceived as intrusive. Missed needs or inappropriate interventions are especially problematic in pervasive settings. Evaluation metrics should account for varying user preferences regarding how much intervention they want from these systems. Addressing these issues requires careful consideration from multiple perspectives within the community.

9.11 Provable and Verifiable Information Retrieval

Provable and verifiable IR approaches in which correctness, fairness, or security can be formally demonstrated and independently audited aim to increase transparency and trust by enabling users or third parties to directly check system behaviour rather than relying solely on claims made by the system itself. As IR systems increasingly rely on AI techniques to meet user needs, there is a risk that these complex models may behave unpredictably or fail in subtle ways. Recent research has shown that formal verification methods can be applied even to neural networks and large AI models. This opens new possibilities for ensuring that AI-powered IR systems remain accountable, helping them meet ethical standards, regulatory requirements, and user expectations while reducing potential harms.

While IR has a strong tradition in evaluation, it has essentially focused on the effectiveness and efficiency of the deployed methods and systems. With the rise of AI-powered methods, new requirements such as fairness, explainability, factuality, and safety have become important. Existing evaluation approaches that rely on test collections are often inadequate or impractical for assessing these broader system properties. A major challenge is developing IR systems that can be formally verified for robustness, transparency, and safe operation while providing auditable records of their decisions and workflows. Creating formal certification processes will help ensure compliance with emerging ethical, legal, and regulatory standards. Addressing these issues is essential to build trust in AI-driven information access systems, especially in sensitive domains like healthcare or finance, where accountability is critical. Provable and verifiable IR supports compliance with the growing number of ethical, legal, and governance frameworks that regulate AI systems worldwide (e.g. UK, EU). It improves safety and security by making IR systems more reliable and less likely to cause harm to users or content producers. By providing auditable and explainable results, it also builds greater trust, accountability, and transparency.

Developing formal verification methods for complex IR systems is challenging. Verifying every aspect of an IR system may not be feasible, so a practical approach is to define general properties and specifications that systems must satisfy and address these. Architectural modifications made for auditing and verification purposes should avoid increasing deployment costs or negatively impacting efficiency and responsiveness. In practice, trade-offs between auditability, cost and performance are likely unavoidable. In addition, as AI methods continue to advance, new regulatory requirements will emerge; this makes it essential to develop processes that can be adapted.

Key Challenges.

- Develop modular IR system architectures that can facilitate third-party verification and certification of its components. Such architectures need to be sufficiently flexible to adapt to new requirements and verification processes.
- Develop tailored model checking methods to verify the intended operation of the IR system and its components, for example, against specified axioms, properties and requirements the IR system should meet.
- Develop workbenches and auditing tools that allow stakeholders to undertake audits of the IR system and its components to assess the quality, the traceability of responses, and the potential harms of AI-powered systems.
- Revisit IR evaluation to include, in addition to effectiveness, standard metrics and auditing benchmarks to evaluate the system's verifiability and compliance with ethical, privacy and regulatory frameworks.

9.12 Quantum Computing for IR

Quantum IR applies ideas and mathematics from quantum mechanics to formulate new models and approaches in IR, regardless of whether these methods are run on actual quantum hardware. In contrast, quantum computing allows algorithms to be executed on real quantum devices. This technology has moved beyond theory and isolated lab work and is becoming increasingly practical and widely accessible for real-world applications, thanks in part to modern development frameworks. Modern IR systems, particularly those using advanced GenAI and ML, require increasing computational power (see Section 2). Quantum computing promises efficiency improvements and could make it possible to solve exponentially complex problems that are not practical or even possible with traditional computing methods.

Key Challenges. Quantum computing applications in IR and recommender systems are still in their early stages. The IR community needs to investigate how existing algorithms can be adapted and executed on quantum devices. This research could lead to faster solutions due to quantum speedups and improved effectiveness by enabling exact solutions instead of relying on heuristics or approximations required by classical computers. Areas such as feature selection, clustering, and instance selection are already suitable for the initial exploration of quantum hardware. Adopting quantum computing also raises new questions about evaluating these solutions in terms of efficiency, complexity, and effectiveness. Since quantum technology is less mature than classical computing, it faces several technical limitations and error sources that must be addressed through proper benchmarking. Evaluation methodologies specific to quantum IR are still lacking, with only a few pioneering efforts underway.

Finally, the perceived high barrier of entry should not discourage researchers from engaging with this field. Recent advances enable algorithm development without in-depth expertise in quantum mechanics; modern frameworks make it accessible for computer scientists familiar with conventional programming techniques.

9.13 Retrieval-Enhanced Machine Learning

The vast majority of ML systems, including LLMs like ChatGPT, LLaMA, and DALL-E, are designed as self-contained systems, with both knowledge and reasoning encoded in model parameters. However, they suffer from several shortcomings:

- these models cannot work effectively for tasks that require knowledge grounding, especially in the case of non-stationary data where new information is actively being produced and tasks that require reconciling conflicting information across a corpus;
- they cannot be applied to long input sequences, or if they can, they carry a significant financial, infrastructure, and computation cost;
- the knowledge in training data is encoded in model parameters, therefore, explanations of their predictions often appeal to abstract and difficult-to-interpret concepts; and
- they do not allow us to value and correct knowledge because they are embedded in their parametric knowledge.

To address these issues, ML systems are increasingly being enhanced with the capability of retrieving knowledge. For example, because a retrieval index is decoupled from model parameters, ML models can access fresh content and generate grounded outputs. They can also partition long input sequences (e.g., long documents or videos) and retrieve the parts that can influence the final predictions. We refer to this research area as retrieval-enhanced ML, including retrieval-augmentation as a special case.

Research on retrieval-enhanced ML is heavily driven from an ML perspective, where the emphasis is on developing predictive models that can leverage retrieval models for prediction effectiveness. This has led to the development of successful retrieval-enhanced ML models, such as Fusion-in-Decoder, RAG, REALM, Guided Transformer, and RETRO. However, most efforts take the retrieval component of retrieval-enhanced ML for granted. This has motivated us to offer a fresh perspective on retrieval-enhanced ML through an IR lens. For example, an IR perspective allows us to frame the retrieval component in retrieval-enhanced ML as a search engine capable of supporting one or more independent predictive models, as opposed to a single predictive model, as is the case in the majority of existing work.

Key Challenges.

- Integrate retrieval modules directly into downstream ML models trained end-to-end.
- Design architectures that write information to memory for later retrieval during inference.
- Developing optimal methods for representing knowledge from a corpus for downstream ML systems.

- Incorporate retrieval at all stages of ML development, including pre-training, to separate memorisation from generalisation. This also helps clarify the distinction between knowledge storage and language modelling in LLMs.
- Create optimisation algorithms that provide richer feedback from the target ML model to the retrieval component.
- Build models that adapt well to new and changing data types, especially under non-stationary conditions.
- Develop methodologies for evaluating retrieval models in the context of retrieval-enhanced ML that go beyond just end-to-end effectiveness.

Acknowledgements: The authors would like to thank Fernando Diaz for his input to this section as he was unable to attend in person.

9.14 Risks of Synthetically Generated Content

The ease and low cost of generating text and multimodal content with GenAI has lowered the barrier to content generation. This content can be valuable when co-authored with humans, enabling people to communicate more easily and effectively despite language or education barriers. However, this content can harm the information and academic ecosystem when employed for profit or manipulation. As one recent example, a computer science conference found synthetic reviewer profiles, submitting auto-generated reviews, on auto-generated publications. Given sufficient collaboration from bad actors, this could lead to an erosion of trust in the review process and pollution of the pool of published papers. Other risks from auto-generated documents include even larger risks of spam, phishing, misformation, and infiltration of the training data for LLMs.

Significant social risks are also associated with AI-generated text, video, and audio (e.g., podcasts, music) content. While these tools offer benefits such as supporting language learning and enabling communication, they also introduce harmful effects. Key concerns include how to define unethical or unacceptable uses of generative technology, how to detect misuse given the speed and scale at which content can be produced, and how difficult it is to distinguish synthetic from human-generated material. For the IR community, such content raises serious challenges in authorship attribution, increases the risk of spreading false or misleading information, and can potentially cause harm to specific groups.

Key Challenges.
Algorithmic research challenges:

- **Watermarking synthetic content:** Develop methods for embedding watermarks into AI-generated text, along with APIs to enable the detection of these marks in documents.
- **Signing human-generated content:** Create technologies that let authors digitally sign their work to prove its human origin.
- **Detecting synthetic content:** Build algorithms capable of distinguishing between human-written and AI-generated documents, since this remains challenging for most people.
- **Labelling synthetic content:** Design metadata standards and systems to clearly label AI-generated text in situations where it is acceptable, enabling consumers to make informed decisions. Early research shows that such labels may influence sharing behaviour.

- **Fact checking at scale:** Advance automated and community-driven methods for verifying claims and identifying misinformation in human and synthetic content.

Social research challenges:

- How can we measure and predict the long-term effects of exposure to AI-generated content through IR systems on individuals, communities, and society?
- How can we ensure that IR systems handling synthetic content remain aligned with ethical, cultural, or community values, particularly in preventing harm or misinformation?
- What methods can identify and address misalignments when IR systems amplify or propagate harmful or misleading synthetic content?
- How can users be meaningfully involved in evaluating and improving IR systems that distribute synthetic content while protecting them from potential negative impacts?

9.15 Semi-Structured Query Languages

Search queries are often preprocessed before retrieval, using steps like stemming, normalisation, and stop-word removal. For LLMs, input preparation can involve creating prompts with instructions or examples to guide the model. RAG systems can make things more complex, sometimes needing rewritten or expanded queries to get better results. In many cases, it helps to be clear about what the user wants, especially when context, like past conversation or location, can make answers more relevant and easier to explain. In addition, various models need different ways of preparing inputs; for example, some require special prompts or tokenisation methods. Furthermore, prompting styles differ; while some models respond best to step-by-step instructions, others work better with statements of intent.

Previous work has introduced database operators and structured query languages that support text-based joins, aggregation, and more expressive search. Systems like TIJAH and XQuery-Fulltext enable structured queries with operators such as join and project, while InQuery and NEXI provide structured retrieval but lack advanced data aggregation. Web search tools like NEAR and quoted phrases demonstrate ways to refine queries. Recent research also explores LLM reasoning with open-ended prompts, though effectiveness depends on prompt structure, as seen in models like T5 that use task-specific prefixes but require precise phrasing for best results.

While research explores prompting "tricks", many lack generalisability. Practical insights, such as user preferences or contextual hints like geolocation, remain underused but can greatly enhance data relevance and user satisfaction. A semi-structured query language can bridge the gap between complex retrieval systems and user-friendly interfaces. By abstracting model intricacies and providing intuitive controls (see Figure 4), this solution can improve relevance, explainability, and adaptability, benefiting both end-users and system developers.

Key Challenges.

- **Specify desired information:** Indicate content focus (e.g., request an abstract vs. product description).
- **Provide feedback:** Mark relevant/irrelevant parts in intermediate retrieval outputs.
- **Separate instructions:** Distinguish between retrieval parameters and LLM generation instructions.

```
{"session": {
    "session_id": "climate_study_001",
        "goal": "Explore climate change impacts through iterative
        refinement.",
        "rounds": [
        {
        "round_id": 1,
        "retrieval": {"query": "climate change environmental impacts",
            "top_k": 5},
        "generation": {"instructions": ["Summarize environmental impacts."],
            "output_format": "paragraph"},
        "feedback": {"user_comments": "Include economic context next."}
        },
        {
        "round_id": 2,
        "retrieval": {"query": "climate change economic impacts",
            "top_k": 3},
        "generation": {"instructions": ["Compare environmental and
            economic impacts."], "output_format": "bullet_points"},
        "feedback": {"user_comments": "Focus on policy implications."}
        },
        {
        "round_id": 3,
        "retrieval": {"query": "climate change policy impacts",
            "top_k": 4},
        "generation": {"instructions": ["Summarize policy impacts,
            integrating prior rounds."], "output_format": "table"}
        }
        ]
  }
}
```

Figure 4. An example of a complex query language that represents instructional feedback over multiple turns for the query "*Retrieve abstracts about climate change that contradict each other, focus on environmental impacts, and provide a rationale before generating the summary.*"

- **Control output easily:** Use intuitive query operators (e.g., `"exact phrase"`, `NEAR`) without needing insider knowledge of model training.
- **Order-aware generation:** Reflect the importance of generation order, supporting internal rationale before conclusions.
- **Adjust stochasticity:** Control hyperparameters (e.g., temperature) to balance determinism and creativity without complexity.

9.16 Vagueness, Uncertainty, and Context

IR is about vagueness and uncertainty in information access.[2] (In this view, database search with precise queries and specific data is a special case of IR.) Vagueness is caused by users' inability to formulate a precise information need, which typically leads to iterative query reformulation. Uncertainty is caused by the system's limited understanding of the query formulation and content due to the limited expressiveness of the underlying representation and the uncertainty of the mapping from the content into this representation. Consequently, traditional IR systems provide a list of answers ranked by decreasing certainty. With LLMs, the methods for addressing these issues have changed. Vagueness may be caused by ambiguous or unspecific terms or fuzzy attribute conditions (e.g., "lightweight laptop with medium-sized screen"). The system can often recognise these situations, then the answer can be structured accordingly, or the user can be asked for clarification. On the other hand, uncertainty is rarely made explicit, as current systems only present (one of the) most likely answer(s) without mentioning the attached uncertainty.

IR should consider context. Traditionally, IR Systems operated context-free (while recommender systems only used users' context as input). Many of today's IR systems maintain a user profile; a Web search systems typically use location and time. Advanced IR systems also consider the user's current situation or task; the best way to achieve this is to integrate the IR component with the application used to perform the task (just like database searches are invoked by application systems connecting to the database system). Thus, IR will become integrated into application systems.

Key Challenges. From the IR concepts described above, the following requirements for future LLM-based IR systems (LIRS) can be derived:

- LIRS must clarify answer uncertainty and provide appropriate interaction possibilities.
- LIRS must provide means for capturing the user's context (instead of only allowing for explicit, lengthy descriptions of this context).
- LIRS should support close integration with applications (as is already happening for other LLM-based services).

10 Conclusion

IR remains a vibrant and essential research area in academia and industry. Addressing people's information needs is a complex, multi-disciplinary challenge, and this report highlights several key research themes within the field. Recent advances in LLMs have accelerated progress, reshaping how we think about search, relevance, and user interaction. These findings are not intended to be exhaustive; many more compelling directions were proposed than could be fully explored during our short time in Torquay. We hope that SWIRL will continue to inspire future research and strategic discussions in this dynamic and rapidly evolving domain.

[2] The 1991 charter of the IR specialist group in the German computer society GI stated "IR deals with vagueness and uncertainty in information systems".

Acknowledgment

This work was developed on the unceded lands of the Wadawurrung People; the Gulidjan and Gadubanud Peoples of the Eastern Maar; and the Woi Wurrung and Boon Wurrung language groups of the Eastern Kulin Nation, in and around what is also known as Torquay and Melbourne, Australia. We acknowledge the Traditional Owners of these lands, their Ancestors and Elders, past and present, and their connection to Country. We respectfully recognise that sovereignty was never ceded and affirm our commitment to truth-telling, recognition, and ethical research collaboration.

A Authors and Affiliations

Editors:

- Johanne R. Trippas, RMIT University, Australia.
- J. Shane Culpepper, The University of Queensland, Australia.

Authors and participants:

- Mohammad Aliannejadi, University of Amsterdam, The Netherlands.
- James Allan, University of Massachusetts Amherst, USA.
- Enrique Amigó, UNED, Spain.
- Jaime Arguello, University of North Carolina at Chapel Hill, USA.
- Leif Azzopardi, Microsoft, Scotland.
- Peter Bailey, Canva, Australia.
- Jamie Callan, Carnegie Mellon University, USA.
- Rob Capra, University of North Carolina at Chapel Hill, USA.
- Nick Craswell, Microsoft, USA.
- Bruce Croft, University of Massachusetts Amherst, USA.
- Jeff Dalton, University of Edinburgh, Scotland.
- Gianluca Demartini, University of Queensland, Australia.
- Laura Dietz, University of New Hampshire, USA.
- Zhicheng Dou, Renmin University of China, China.
- Carsten Eickhoff, University of Tübingen, USA.
- Michael Ekstrand, Drexel University, USA.
- Nicola Ferro, University of Padua, Italy.
- Norbert Fuhr, University of Duisburg-Essen, Germany.
- Dorota Glowacka, University of Helsinki, Finland.
- Faegheh Hasibi, Radboud University, The Netherlands.
- Danula Hettiachchi, RMIT University, Australia.
- Rosie Jones, Spotify, USA.
- Jaap Kamps, University of Amsterdam, The Netherlands.
- Noriko Kando, National Institute of Informatics, Japan.
- Sarvnaz Karimi, CSIRO, Australia.
- Makoto P Kato, University of Tsukuba, Japan.
- Bevan Koopman, CSIRO and University of Queensland, Australia.

- Yiqun Liu, Tsinghua University, China.
- Chenglong Ma, RMIT University, Australia.
- Joel Mackenzie, The University of Queensland, Australia.
- Maria Maistro, University of Copenhagen, Denmark.
- Jiaxin Mao, Renmin University of China, China.
- Dana McKay, RMIT University, Australia.
- Bhaskar Mitra, Microsoft, Canada.
- Stefano Mizzaro, University of Udine, Italy.
- Alistair Moffat, The University of Melbourne, Australia.
- Josiane Mothe, University of Toulouse, France.
- Iadh Ounis, University of Glasgow, Scotland.
- Lida Rashidi, RMIT University, Australia.
- Yongli Ren, RMIT University, Australia.
- Mark Sanderson, RMIT University, Australia.
- Rodrygo Santos, Universidade Federal de Minas Gerais, Brazil.
- Falk Scholer, RMIT University, Australia.
- Chirag Shah, University of Washington, USA.
- Laurianne Sitbon, Queensland University of Technology, Australia.
- Ian Soboroff, NIST, USA.
- Damiano Spina, RMIT University, Australia.
- Paul Thomas, Microsoft, Australia.
- Julián Urbano, Delft University of Technology, The Netherlands.
- Arjen de Vries, Radboud University, The Netherlands.
- Ryen White, Microsoft, USA.
- Abby Yuan, The University of Melbourne, Australia.
- Hamed Zamani, University of Massachusetts Amherst, USA.
- Oleg Zendel, RMIT University, Australia.
- Min Zhang, Tsinghua University, China.
- Justin Zobel, The University of Melbourne, Australia.
- Shengyao Zhuang, CSIRO, Australia.
- Guido Zuccon, The University of Queensland, Australia.

References

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 the Chinese IR community. *AI Open*, 4:80–90, 2023.

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. ISSN 0163-5840. doi: 10.1145/2215676.2215678. URL https://doi.org/10.1145/2215676.2215678.

James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. Future of Information Retrieval research in the age of generative AI. *arXiv preprint arXiv:2412.02043*, 2024.

J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. Research frontiers in information retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52(1):34–90, August 2018. ISSN 0163-5840. doi: 10.1145/3274784.3274788. URL https://doi.org/10.1145/3274784.3274788.

Alistair Moffat, Justin Zobel, and David Hawking. Recommended reading for IR research students. *SIGIR Forum*, 39(2):3–14, December 2005. ISSN 0163-5840. doi: 10.1145/1113343.1113344. URL https://doi.org/10.1145/1113343.1113344.