Report from Dagstuhl Seminar 25391

Retrieval-Augmented Generation—The Future of Search?

Edited by

Matthias Hagen[1], Josiane Mothe[2], Smaranda Muresan[3], Martin Potthast[4], Min Zhang[5], and Benno Stein[6]

1 **Friedrich-Schiller-Universität Jena, DE,** matthias.hagen@uni-jena.de
2 **Toulouse University, FR,** josiane.mothe@irit.fr
3 **Barnard College, Columbia University - New York, US,** smara@columbia.edu
4 **University of Kassel, hessian.AI, ScaDS.AI, DE,** martin.potthast@uni-kassel.de
5 **Tsinghua University - Beijing, CN,** z-m@tsinghua.edu.cn
6 **Bauhaus-Universität Weimar, DE,** benno.stein@uni-weimar.de

Abstract

Dagstuhl Seminar 25391 "Retrieval-Augmented Generation—The Future of Search?" was held in the week of September 21–26, 2025. Fourty researchers, most of whom came from the fields of information retrieval and web search as well as natural language processing, were invited to share the latest developments in the area of retrieval-augmented generation and discuss its research agenda and future directions. The 5-day program of the seminar consisted of four introductory and background sessions, two short talks sessions about technology and demos, one industry talk session, one afternoon hackathon, and nine working groups and reporting sessions. The seminar also had three social events during the program. This report provides the executive summary, overview of invited talks, and findings from the five working groups which cover the potential and limitations, information behavior and result presentation, the system side, societal and ethical aspects, and the evaluation of retrieval-augmented generation. The ideas and findings presented in this report should serve as one of the main sources for diverse research programs on retrieval-augmented generation.

Seminar September 21–26, 2025 – http://www.dagstuhl.de/25391
2012 ACM Subject Classification Information Systems; Artificial Intelligence

Retrieval-Augmented Generation—The Future of Search?, *Dagstuhl Reports*, pp. 1–94
Editors: Matthias Hagen, Josiane Mothe, Smaranda Muresan, Martin Potthast, Min Zhang, Benno Stein
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Matthias Hagen (Friedrich-Schiller-Universität Jena, DE)
Josiane Mothe (Toulouse University, FR)
Smaranda Muresan (Barnard College, Columbia University - New York, US)
Martin Potthast (University of Kassel, hessian.AI, and ScaDS.AI - DE)
Min Zhang (Tsinghua University - Beiing, CN)
Benno Stein (Bauhaus-Universität Weimar, DE)

Background and Motivation

Retrieval-augmented generation (RAG) has proven effective in conditioning the output of large language models (LLMs) on relevant documents and for grounding LLM-generated statements, this way combatting the so-called hallucination or confabulation problem. Basically, RAG combines (1) a retrieval phase where a search system identifies relevant documents for a user prompt and (2) a generation phase, where an LLM synthesizes a tailored answer, probably linking to the retrieved sources.

RAG challenges "classical" retrieval technology and has the potential to revolutionize information-seeking behavior overall by reducing the searcher's effort to extract the desired information from individual search results. The revolution becomes evident, among others, in a change in the design of a search engine results page (SERP): Instead of presenting the proverbial list of "ten blue links", the list SERP, a generated text with references is shown, the text SERP. The first public prototypes of this kind were You.com's You Chat and the now discontinued Neeva AI, followed by Microsoft's Bing Chat, Google's Bard, Perplexity.ai, and Baidu's Ernie. However, many unsolved problems and relevant research questions still lurk under the hood (i.e., the user interface).

The proposed Dagstuhl Seminar will focus on the expectations, the promises, the potential, and the limits of integrating RAG in information retrieval (IR). Relevant questions include:

- Will we ever search again?
- Will RAG bias retrieval results?
- Is RAG more than fact checking for conversational IR?
- How can we measure the effectiveness of RAG-based systems?
- How can we keep RAG-based systems transparent and accountable?

To work on these and related questions, the Dagstuhl seminar will bring together experts from the fields of information retrieval, natural language processing, and generative AI who have academic, non-profit (e.g., MetaGer or the Open Search Foundation), or industrial backgrounds (e.g., You or Microsoft).

Seminar Program

A 5-day program of the seminar consisted of five introductory and background sessions, one perspectives talk session, one industry talk session, and nine breakout discussion and reporting sessions. The program also had three social events and is available online.[1]

[1] https://www.dagstuhl.de/schedules/25391.pdf

Pre-Seminar Activities

Prior to the seminar, participants were asked to provide inputs to the following questions and request:

1. Will RAG replace ranked search for end users?
2. Please list, from the perspective of your research interests, important open questions or challenges in RAG.
3. What are the three papers a PhD student in RAG should read and why?

The first question has been answered by 30 out of the 40 participants. The answers were almost evenly distributed between "No" with 16 votes and "Yes" with 13 votes. In addition to asking this question, we asked participants to give a reason for their vote. These answers turned out to be nuanced, including various forms of hedging one way of the other. A list of the arguments made is provided in Section 5 of this report.

From the survey, the following topics initially emerged as interests of participants. Many of these topics were discussed at length in the seminar.

- Can the methodologies that underpin RAG solve the enterprise search problems that sparse retrieval struggled to deal with?
- How can biases in the responses of RAG systems be detected?
- How to effectively inject external knowledge into LLMs?
- How to conduct RAG efficiently and dynamically?
- What is the analog of the Cranfield-style evaluation of ranked retrieval for RAG?
- What should the role and prominence of citations be in RAG outputs?
- How does RAG influence information behavior and how does it affect relevance feedback?
- How should search engines be designed that are used by RAG agents?
- What role does reasoning play in RAG?
- How can RAG systems be trained and how can they aid in training generative AIs?
- How can RAG systems notice and express uncertainty in their response?

Another outcome of the above pre-seminar questions was the compilation of a list of recommended reading to gain a solid understanding of topics and technologies related to retrieval-augmented generation. The reading list is provided in Section 6 of this report.

Invited Talks

One of the main goals and challenges of this seminar was to bring a broad range of researchers together to discuss retrieval-augmented generation, which required to establish common terminologies among participants. Therefore, we had a series of 26 invited talks throughout the seminar program to facilitate the understanding and discussion of retrieval-augmented generation and its potential enabling technologies. Section 3 contains the abstracts of all talks.

Working Groups

In the afternoon of Day 2, initial working groups were formed based on the inputs to the pre-seminar questionnaires, introductory, and background talks, and discussions among participants. Eventually, the following five groups were formed:

- Potential & Limitations with Respect to Cognitive Tasks
- Information Behavior and Result Presentation in RAG
- Retrieval-Augmented Generation: The System's Perspective
- Societal and Ethical Motivations for Inverting RAG to GAR
- An Unexamined RAG Is Not Worth Interrogating

We have summarized the working groups' outcomes in the following. Please refer to the main part of this report for the full description of the findings.

Potential & Limitations with Respect to Cognitive Tasks

The aim of this group was to provide an overview of the cognitive tasks in which a user can be supported by retrieval-augmented generation (RAG) systems. These cognitive tasks were organized in a taxonomy and linked to Bloom's taxonomy. With Bloom's taxonomy in mind, it became obvious that RAG is the superior technology in comparison to ranked search for information needs for which the solution requires synthesis (i.e., activates all layers in Bloom's taxonomy). However, if a user outsources these tasks to RAG systems frequently, there are risks of cognitive decline for the user since cognitively demanding synthesis tasks are rarely performed manually. These risks were outlined by this group and scaffolding strategies to mitigate them were proposed.

Information Behavior and Result Presentation in RAG

This group examined how RAG systems can cater to a diverse range of information behaviors and tailor the result presentation to the requirements of different users and tasks. They discussed how RAG systems allow for more natural interactions in multiple modalities and the formulation of complex queries that are contextualized by the rich history of previous interactions.

The group emphasized the need for adaptive, transparent, and inclusive RAG systems that accommodate diverse users and tasks - from factual retrieval to creative generation. They analyzed interaction dimensions such as user versus system initiative, information complexity, and human–machine collaboration. Their discussions resulted in open research questions around provenance tracking, adaptive presentation, and fostering user engagement while counteracting cognitive offloading.

Retrieval-Augmented Generation: The System's Perspective

This group took a systems-level view of RAG, examining its architecture and contrasting today's naive setup with an ideal RAG system. First, they highlighted that retrieval models were originally built for human users, informing evaluation methods and result formats that may not be suited for LLMs as users. Second, they emphasized the many sources of uncertainty in RAG, such as retrieval, model reasoning, and data quality, and the need for better ways to detect and communicate this uncertainty. Third, they explored efficiency–effectiveness trade-offs, arguing that future systems should dynamically allocate computation based on query complexity. Finally, the group discussed the advantages of federated RAG systems, such as unified access to open and proprietary sources, as well as challenges that this kind of architecture would face.

Societal and Ethical Motivations for Inverting RAG to GAR

This group advocated for a shift from retrieval-augmented generation (RAG) to Generation-Augmented Retrieval (GAR) to develop next-generation information access tools. This framing prioritizes information retrieval and the design of transparent, ethical, and sustainable information access systems that encourage active user engagement with information sources and diverse knowledge ecosystems. First, the group discussed the intersection of knowledge, ethics, and human rights, and its implications for RAG vs. GAR. Next, they offered perspectives on several (though not exhaustive) sociotechnical issues for GAR: scholarly communication, user cognition, emotional and mental well-being, democracy and political discourse, and language and culture. Each section references existing challenges related to the rapid and widespread use of generative AI to encourage critical and informed thinking about GAR development for individuals, organizations, and society as a whole. Throughout, the group proposed considerations for information retrieval researchers to create information access systems, learn from user studies, and foster interdisciplinary partnerships.

An Unexamined RAG Is Not Worth Interrogating

This group focused on how one can determine in which scenarios RAG systems work, where they fail, and how one can identify which RAG system is suitable for which scenario. The motivation of the group was that paradigm shifts in information access technology might also require paradigm shifts for corresponding evaluations. It therefore started by hypothesizing which evaluation properties might change between mature information retrieval evaluation methods and the methods applied in RAG systems. The group recognized that RAG evaluation is currently in an exploratory stage where many different evaluation ideas are being explored. The group identified research gaps and proposed directions for future research. Many parts of the discussion evolved around the idea of enabling in-depth analysis of RAG responses by multiple experts. This activity motivated the group to brainstorm how concepts from the Talmud can be transferred to RAG evaluations and labeling.

Conclusions

Leading researchers from diverse domains in academia and industry investigated the essence, attributes, architecture, applications, challenges, and opportunities of retrieval-augmented generation in the seminar. One clear signal from the seminar is that research opportunities to advance retrieval-augmented generation are available to many areas and collaboration in an interdisciplinary community is essential to achieve this goal. This report should serve as one of the main sources to facilitate such diverse research programs on retrieval-augmented generation.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups

3 Overview of Talks

3.1 Legal Retrieval and Augmented Generation

Qingyao Ai (Tsinghua University - Beijing, CN)

The rapid progress of large language models (LLMs) has opened new opportunities for legal artificial intelligence, yet the legal domain presents unique challenges that require specialized solutions. This talk explores legal retrieval-augmented generation (Legal RAG) as a framework to address these challenges, focusing on data, augmentation, evaluation, and applications. Legal case documents are structurally complex and lengthy, with relevance definitions that differ fundamentally from general domains. To manage extreme input lengths and heterogeneous sources—such as statutes, case documents, and legal essays—RAG methods should dynamically retrieve and integrate multi-source knowledge for reasoning tasks. We highlight applications across legal judgment prediction, legal document writing, and court simulation. Evaluation remains difficult as it requires significant legal expertise and is highly subjective in many cases. We introduce our initial efforts on building taxonomies and benchmarks for legal RAG and LLMs, and hope this would help people develop better legal models in future.

3.2 Parametric Retrieval-Augmented Generation

Qingyao Ai (Tsinghua University - Beijing, CN)

Retrieval-augmented generation (RAG) has emerged as a promising solution to enhance the reliability of large language models (LLMs) with external knowledge. Existing RAG methods share a common strategy for knowledge injection: they place the retrieved documents into the input context of the LLM, which we refer to as the in-context knowledge injection method. While this approach is simple and often effective, it has inherent limitations. Firstly, increasing the context length and number of relevant documents can lead to higher computational overhead and degraded performance, especially in complex reasoning tasks. More importantly, in-context knowledge injection operates primarily at the input level, but LLMs store their internal knowledge in their parameters. This gap fundamentally limits the capacity of in-context methods. To this end, we introduce Parametric RAG, a new RAG paradigm that integrates external knowledge directly into the feed-forward networks of an LLM through document parameterization. This approach not only reduces online computational costs by shortening the input context length, but also deepens the integration of external knowledge by enabling LLMs to utilize it in the same way as internal parametric knowledge. Experimental results demonstrate that Parametric RAG substantially enhances the effectiveness and efficiency of knowledge augmentation in LLMs. Also, it can be combined with in-context RAG methods to achieve even better performance.

3.3 Interactions with RAG

Mohammad Aliannejadi (University of Amsterdam, NL)

In this talk, I present the current challenges of generative and RAG systems in terms of user interaction. How do the existing user interaction and information need levels generalize to RAG systems and what are the implications of the new chat-based interfaces on user experience? In particular, how do the changes affect the click behavior of users and what are the risks of that in terms of trust and factuality of the results?

3.4 An Industry Perspective on RAG

Sophia Althammer (Cohere - München, DE)

This talk gives an overview of products with RAG (retrieval-augmented generation) and Agents as well as outlining important aspects for enterprise customers. We define agents and give an overview how the Command-A model was trained with respect to agentic capabilities. We also touch on evaluation benchmarks and general progress of agents and outline possible open research questions on agents.

3.5 RAGE: How to Evaluate RAG Systems

Laura Dietz (University of New Hampshire - Durham, US)

Retrieval-augmented generation (RAG) has become a central paradigm for knowledge-intensive applications, yet its evaluation remains a persistent challenge. Traditional relevance-based evaluation, grounded in human judgments over static documents, does not translate directly to contexts where each query produces a novel, free-form response. In response, researchers have increasingly employed large language models (LLMs) as judges, using methods such as direct prompting, pairwise preference comparison, multi-criteria prompting, nugget-based evaluation, and multi-step frameworks. These approaches offer scalability and fine-grained assessment, but they also introduce new risks.

A prime threat is circularity, in which systems and evaluations rely on the same or overlapping models, producing results that appear plausible yet fail to align with human judgments. The talk demonstrates that direct prompting methods are particularly vulnerable to this effect, undermining their reliability under meta-evaluation. Nugget-based LLM judges, by contrast, promise more interpretable and granular assessments, but their resilience is not guaranteed: if evaluation “secrets” such as gold nuggets or rubric structures can be anticipated or guessed by systems, then even nugget-based evaluation may be compromised.

This talk presents an investigation into often overlooked factors that negatively impact evaluation methods for RAG systems and the meta-evaluation of those methods. It argues that evaluation is no longer a passive measure of progress but an active force shaping system design, with feedback loops that must be understood to avoid distorted conclusions.

3.6 RAG 4 Med

Carsten Eickhoff (Universität Tübingen, DE)

The modern healthcare system and its various actors face unprecedented staffing shortages and tightening economic constraints. Data driven methods, including generative AI applications offer a promising means of automating, and enhancing many steps in the clinical pipeline. In this talk, we will briefly discuss the opportunities and challenges that these applications face and which role retrieval augmented generation can play in this high-stakes environment.

3.7 Core IR Concepts and RAG

Norbert Fuhr (Universität Duisburg-Essen, DE)

Information retrieval is about vagueness, uncertainty and context in information access. Vagueness is caused by the fact that users cannot give a precise specification of their information need, thus using vague query conditions and iterative query formulation. While the latter is addressed by conversational IR, LLM-based systems usually choose the most probable meaning for vague query terms. Uncertainty is due to the system's uncertain knowledge about the user's information need and the database objects. Traditional IR system use ranking for dealing with uncertainty, but more advanced strategies in RAG are an open issue. Consideration of context requires an IR system to take into account both the user and the situation when answering a query. Currently, RAG users enrich their queries with lengthy descriptions of the context. Overall, future RAG systems should explicitly address vagueness in queries, clarify answer uncertainty and integrate means for capturing user context automatically.

3.8 MetaRAG: Learning About RAG from a RAG System

Marcel Gohsen (Bauhaus-Universität Weimar, DE)

Exploring the landscape of papers about retrieval-augmented generation (RAG) can be a challenge due to the sheer volume of publications. Specifically for the Dagstuhl seminar on RAG, we developed "MetaRAG", a conversational search system that uses RAG to facilitate navigation through the literature. MetaRAG can answer open-ended question or summarize a specific publication, retrieving from a collection of over 3,000 papers.

3.9 OpenWebSearch.eu: An Open Scaleable Infrastructure for Web Search & RAG

Michael Granitzer (Universität Passau, DE)

OpenWebSearch.eu project aims to develop an open European infrastructure for web search. The project aspires to contribute to Europe's digital sovereignty and help promote an open human-centred search engine market.

OpenWebSearch.eu is designing the core of the European Open Web Index, based on open source software and deployed across various high performance computing centres in Europe. The Open Web Index is particularly crucial for the provision of state-of-the-art web search services and for European innovations, such as AI/large language models.

We build an infrastructure to establish and maintain relevant web-data sets at Petabyte-scale, including raw web-data, web-index and large multilingual text corpora as well as multimodal data sets at the various EuroHPC centres: ready-to-use by researchers, start-ups, innovators and industry to build value oriented, trustworthy AI/LLM and search solutions in and for Europe. We are confident that sharing these data sets across Europe by hosting them very close to or directly at the HPC facilities would be the most effective way.

An Open Web Index will also provide new opportunities in developing decentralised RAG systems by providing relevant web-data for bootstrapping topic specific search engines.

3.10 Uncertainty Quantification for RAG

Faegheh Hasibi (Radboud University Nijmegen, NL)

Uncertainty Quantification (UQ) provides methods to estimate an LLM's confidence in its outputs and helps users assess the reliability of its responses. Ideally, an effective UE method would assign low uncertainty to correct answers and high uncertainty to incorrect ones. However, existing UE methods have not been thoroughly studied in the context of retrieval-augmented generation (RAG). Recent work shows that current approaches often fail to reliably estimate response correctness in the simple retrieve-then-generate paradigm of RAG. In this talk, we will explore the applications of UE in information access systems and discuss the challenges of uncertainty estimation in more complex RAG settings, where generation unfolds through multiple reasoning and retrieval steps.

3.11 Ads in RAG: Can We Block Promotional Text in LLM Responses?

Sebastian Heineking (Universität Leipzig, DE)

Joint work of Heineking, Sebastian; Zelch, Ines; Bevendorff, Janek; Stein, Benno; Hagen, Matthias; Potthast, Martin

Main reference Schmidt, Sebastian; Zelch, Ines; Bevendorff, Janek; Stein, Benno; Hagen, Matthias; Potthast, Martin, Detecting Generated Native Ads in Conversational Search, Companion Proceedings of the ACM Web

Conference 2024, pp. 722–725.
URL https://doi.org/10.1145/3589335.3651489

Due to its prevalence in classic search engines, advertising is a likely business model for retrieval-augmented generation (RAG). As one example, researchers from Google outlined an auction mechanism for individual tokens that illustrates how LLM-based advertising could be implemented [50]. In contrast to conventional advertising, LLMs can tailor advertisements to the user's preferences and current information need, and blend them with the rest of the generated text, making them difficult to detect. This talk gives an overview on our ongoing research on blocking advertisements in LLM responses [87, 141].

3.12 Two Things You Must Know About the G from RAG

Djoerd Hiemstra (Radboud University Nijmegen, NL)

In this short presentation I will discuss: 1) why chatbots and retrieval-augmented generation (RAG) are so irresistible to people, and 2) the energy consumption of text generation approaches based on large language models. I will argue that we MUST NOT assign human traits to chatbots and RAG system in our research papers, and that we MUST estimate the energy consumption of our research.

3.13 Rankify Toolkit for Retrieval, Re-Ranking, RAG & RankArena

Adam Jatowt (Universität Innsbruck, AT)

In this talk I have introduced Rankify toolkit [2] and RankArena comparison and demonstration framework [1] for fostering research on RAG and reranking algorithms.

Rankify is a modular open-source toolkit designed to unify retrieval, re-ranking, and RAG within a cohesive framework. Rankify supports a wide range of retrieval techniques, including dense and sparse retrievers, while incorporating state-of-the-art reranking models to enhance retrieval quality. Additionally, Rankify includes a collection of pre-retrieved datasets to facilitate benchmarking, available at Huggingface. As a unified and lightweight framework, Rankify allows researchers and practitioners to advance retrieval and reranking methodologies while ensuring consistency, scalability, and ease of use. Rankify is available at `https://github.com/DataScienceUIBK/rankify`.

RankArena is a unified platform for comparing and analysing the performance of retrieval pipelines, rerankers, and RAG systems using structured human and LLM-based feedback as well as for collecting such feedback. RankArena supports multiple evaluation modes: direct reranking visualisation, blind pairwise comparisons with human or LLM voting, supervised manual document annotation, and end-to-end RAG answer quality assessment. It captures fine-grained relevance feedback through both pairwise preferences and full-list annotations, along with auxiliary metadata such as movement metrics, annotation time, and quality ratings. The platform also integrates LLM-as-a-judge evaluation, enabling comparison between model-generated rankings and human ground truth annotations. All interactions are stored as structured evaluation datasets that can be used to train rerankers, reward models, judgment agents, or retrieval strategy selectors. Our platform is publicly available at https://rankarena.ngrok.io/, and the Demo video is provided at https://youtu.be/jIYAP4PaSSI.

3.14 Hint-Based Interaction

Adam Jatowt (Universität Innsbruck, AT)

Chatbots have rapidly become integrated into everyday life, offering instant, human-like answers to a wide range of questions. While this technology improves information access, it can also weaken cognitive skills by encouraging shallow processing, contributing to information overload, and fostering over-reliance on automated reasoning. Two features of hint-based interaction make it a promising alternative. First, hints act as scaffolds, guiding users toward answers without revealing them outright. Second, hints are adaptable, meaning they can be tailored to different purposes, question types, and user needs. We discuss the idea of automatic hint generation, the criteria of hint quality estimation and we demonstrate several datasets of hints, as well as introduce the HintEval framework (https://github.com/DataScienceUIBK/HintEval) designed for supporting hint-focused research.

3.15 User Simulation for Generative IR Systems: GenIRSim

Johannes Kiesel (GESIS - Leibniz Institute for the Social Sciences, DE)

Joint work of Kiesel, Johannes; Gohsen, Marcel; Mirzakhmedova, Nailia; Hagen, Matthias; Stein, Benno
Main reference Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. Who Will Evaluate the Evaluators? Exploring the Gen-IR User Simulation Space. In *15th International Conference of the CLEF Association (CLEF 2024)*, pages 166–171, 2024.

The reliable and repeatable evaluation of interactive, conversational, or generative IR systems is an ongoing research topic in the field of retrieval evaluation. One proposed solution is to fully automate evaluation through simulated user behavior and automated relevance judgments. Still, simulation frameworks were technically quite complex and have not been widely adopted. Recently, however, easy access to large language models has drastically lowered the hurdles for both user behavior simulation and automated judgments. We therefore argue that it is high time to investigate how simulation-based evaluation setups should be evaluated themselves. We present GenIRSim, a flexible and easy-to-use simulation and evaluation framework for generative IR.

3.16 The Turing Game

Johannes Kiesel (GESIS - Leibniz Institute for the Social Sciences, DE)
Benno Stein (Bauhaus-Universität Weimar, DE)

Developments in generative AI (ChatGPT, Stable Diffusion, etc.) are redefining the boundaries between humans and machines. However, current debates on this topic primarily revolve around questions of performance: Do machines write better than humans? Are their images more appealing or even more creative than human art works? We argue that the focus on performance overlooks one of the most pressing questions of current development, namely: Do humans accept "intelligent machines" as members of their community?

To answer this question, we are developing the "Turing Game", based on the classic Turing Test [169]. While the Turing Test was conceived as an "imitation game" and made a ground-breaking contribution to the question of whether machines can think, the Turing Game serves to ask how hybrid communities with human and machine members can function.

3.17 Neural and LLM Retrieval

Sean MacAvaney (University of Glasgow, GB)

How do you go from a strong LLM-based relevance model to a complete search engine? This talk gives a brief introduction to the families of retrieval techniques for LLMs.

3.18 Sociotechnical Implications of RAG for Information Access

Bhaskar Mitra (Independent Researcher, Tiohtià:ke/Montréal, CA)

Robust access to trustworthy information is a critical need for society with implications for knowledge production, public health education, and promoting informed citizenry in democratic societies. Generative AI technologies with retrieval-augmentation may enable new ways to access information and improve effectiveness of existing information retrieval systems, but we are only starting to understand and grapple with their long-term social implications. In this talk, we discuss some of the systemic risks of employing generative AI and RAG in the context of information access that should critically inform future research and development in this area.

3.19 Beyond English: Cultural and Linguistic Challenges in RAG Systems

Josiane Mothe (Toulouse University, FR)

This talk probes how retrieval-augmented generation (RAG) handles culture and language. A simple demo—“help me prepare a report on turkey”—shows smart disambiguation but also unexpected code-switching and clumsy French, prompting questions about language choice on the different steps of RAG and what sources get prioritized. Source selection -which is key- shifts with language and locale. Truth and facts are often filtered by context. While RAG offers new opportunities it also come with threats that are explored.

3.20 The Jewish Talmud as the Past (or Beginning?) of RAG

Birte Platow (TU Dresden, ScaDS.AI, DE)

From a humanities perspective, RAG systems raise fundamental questions about (new) knowledge systems: What is the origin and authority of a text? How do we deal with discursive/contradictory statements? To what extent and why do we want to recognize it as “true”? And which interpretations do we want to establish? We have been familiar with these and other questions relating to the handling of text for thousands of years, especially when it comes to “sacred texts.” The Jewish Talmud deals with this in depth and can perhaps be interpreted as an ancient precursor to RAG, posing challenging questions for AI-generated texts and knowledge systems.

3.21 RAG Evaluation

Mark Sanderson (RMIT University – Melbourne, AU)

This talk provides an overview of the two main areas of new research in the evaluation of RAG systems. For the first area, I will describe the use of LLMs to change the evaluation of the retrieval component of a RAG system. Here, I talk about the widely discussed use of LLMs for relevance assessment and the growing interest of using LLMs to simulate other critical aspects of retrieval evaluation, such as queries and document collections. For the second area, I will discuss the evaluation of RAG itself, in particular, I will highlight how RAG evaluation challenges the IR community to tackle aspects of evaluation it has gotten away with ignoring in the past.

3.22 Multimodal LLMs and RAG

Alan Smeaton (Dublin City University, IE)

This talk examines the differences between RAG as used in text-based LLMs, and RAG as it is (not yet) available in mutimodal LLMs. It concludes by demonstrating that text-based RAG has a large advantage over MM-RAG because of the strong heritage and experience of developing multiple IR techniques over recent decades whereas multimedia IR is in it's comparative infancy but at least now it has a roadmap for progress.

3.23 Adapting RAG to Users

Damiano Spina (RMIT University – Melbourne, AU)
Johanne Trippas (RMIT University – Melbourne, AU)

Retrieval-augmented generation (RAG) has proven effective in grounding the outputs of large language models (LLMs) with evidence from retrieved passages, thereby reducing errors or "hallucinations" in responses. Beyond its technical promise, RAG represents a new information-seeking paradigm that reshapes how users interact with systems to satisfy their information needs. In this talk, we argue that RAG – and, more broadly, generative information retrieval (GenIR) – creates opportunities to revisit long-standing concepts and methodologies in information retrieval. We explore how this shift not only introduces new types of information needs and tasks, but also new kinds of "users," including LLMs themselves acting as searchers. This makes it a particularly exciting time to be working on interactive IR and evaluation, as the field expands to address these emerging challenges and opportunities.

3.24 RAG Foundations and Models

Arjen P. de Vries (Radboud University Nijmegen, NL)

Joint work of Su, Weihang; Ai, Qingyao; Zhan, Jingtao; Dong, Qian; Liu, Yiqun; de Vries, Arjen P.

This talk gives an introduction to retrieval-augmented generation (RAG) based on the tutorial 'Dynamic and Parametric Retrieval-Augmented Generation' that Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong and Yiqun Liu gave at SIGIR 2025.

3.25 Advances in LLM Rankers

Guido Zuccon (University of Queensland - Brisbane, AU)

Generative large language models (LLMs) like Gemini, GPT, and Llama are transforming information retrieval, enabling new and more effective approaches to document retrieval and ranking. The switch from the previous generation pre-trained language models backbones (e.g., BERT, T5) to the new generative LLMs backbones has required the field to adapt training processes; it also has provided unprecedented capabilities and opportunities, stimulating research into zero-shot approaches, reasoning approaches, reinforcement learning based training, and multilingual and multimodal applications. In this talk I provide an overview of LLM-based rankers, covering fundamental architectures and open challenges and research directions.

3.26 Do LLMs Search Differently?

Guido Zuccon (University of Queensland - Brisbane, AU)

Search has traditionally been defined as the interaction between a user and an information retrieval (IR) system to satisfy an information need. From the outset, IR evaluation has assumed that system utility lies in serving human users, and that this utility can be approximated by measuring relevance. Consequently, retrieval models and evaluation measures have been designed around human-centred notions of relevance, and search systems have been optimised accordingly. These same search systems are now central to generative AI, particularly in retrieval-augmented generation (RAG), where a retriever supplements the input of a large language model (generator) with external documents. RAG has enabled state-of-the-art performance on knowledge-intensive tasks, improved attribution of answers to sources, and allowed models to in-corporate new knowledge without retraining. Yet, retrieval in RAG still relies on methods optimised for humans rather than for generators. Human-oriented practices such as ranking documents by semantic relevance reflect assumptions about how people search: that they read results top to bottom, expect the most relevant item first, and stop once their need is satisfied. These assumptions are embedded in evaluation measures such as normalised discounted cumulative gain (nDCG), which reward systems for placing relevant documents higher in the list. Yet empirical studies reveal that such ranked lists can impair generators. A model may prefer relevant documents at the end of the list, or even perform better when non-relevant documents appear first. These findings raise a deeper question: what principles should guide the design of retrievers for generators?

4 Working Groups

4.1 Potential & Limitations with Respect to Cognitive Tasks

Liesbeth Allein (KU Leuven, BE)
Sophia Althammer (Cohere, DE)
Nolwenn Bernard (TH Köln, DE)
Marcel Gohsen (Bauhaus-Universität Weimar, DE)
Adam Jatowt (Universität Innsbruck, AT)
Abhinav Joshi (Indian Institute of Technology Kanpur, IN)
Smaranda Muresan (Barnard College, Columbia University, US)
Jian-Yun Nie (University of Montreal, CA)
Benno Stein (Bauhaus-Universität Weimar, DE)

The working group on cognitive tasks and RAG focused on benefits, potentials as well as limitations and challenges of RAG systems with respect to user cognition and cognitive tasks that users complete. The section first compares the different tasks that rank-based retrieval and RAG systems help to solve based on Bloom's taxonomy of learning. It then discusses a taxonomy of different tasks that can be completed with RAG and that are based on synthesizing information to create a solution object. In the remaining part of the section, we discuss the concepts of cognitive offloading and the risks of cognitive decline, as well as propose several future research directions.

4.1.1 Introduction and Motivation

Retrieval-augmented generation (RAG) systems have shown great potential for assisting and potentially automating a manifold of cognitive tasks for humans. Products powered with RAG systems like Google with generative AI summaries [65], ChatGPT Agent [125], ChatGPT DeepResearch [126] or Claude Code [11] are already changing how people interact with the web, how they solve implementation problems, or how they plan trips [72, 124].

RAG systems augment traditional ad-hoc retrieval systems by giving a large language model (LLM) access to different retrieval tools. In order to fulfill a user's request, the model can decide which and how to call the retrieval tools and the system can give back a generated response to the user or do actions directly with the tools given to the model.

Users of RAG systems originate from the whole population and cover children, adults, and elderly. Those different groups all come with different cognitive abilities and use RAG systems with different tasks and goals in mind. Each of these tasks cause different levels of cognitive demand for the user. An example of a task with higher cognitive demand is acquiring a new skill as it goes beyond one-shot information access and involves different aspects of information and practice. A task with lower cognitive demand can be as simple as retrieval of a cooking recipe.

The cognitive demand for the user is interdependent on the demand a human would be exposed to if he would replace the LLM in a RAG system. For example, retrieving (or synthesizing) a cooking recipe causes little cognitive load for a user since he only needs to understand and remember details of the recipe, while a human in the shoes of a RAG system would need to find and judge relevant recipes and extract important information which would cause a high cognitive demand. Vice versa, acquiring a new skill cause high

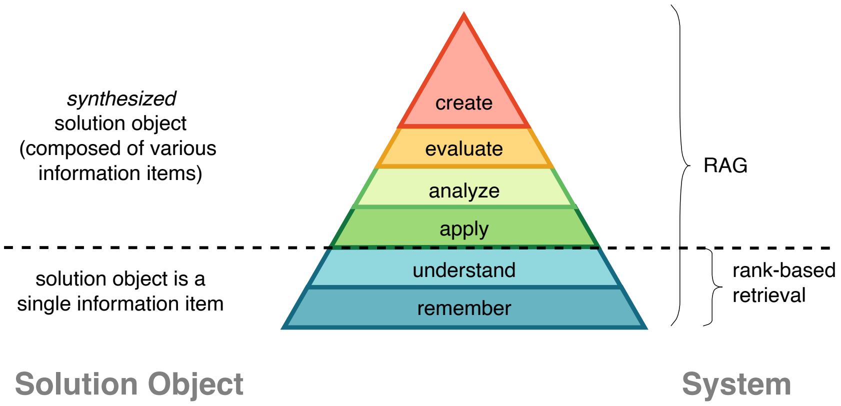


Figure 1 Bloom's revised taxonomy of learning [92] can be used to demonstrate and discuss the complexity of tasks that can be tackled by a RAG system and a rank-based retrieval system, respectively. While RAG systems are capable of synthesizing complex solution objects to tackle cognitive demanding tasks, traditional rank-based systems focus on searching and finding single information items only, so the solution object already needs to exist.

cognitive demand for the user, however, providing teaching materials to the user can be seen as a task of lower cognitive demand. In chapter 4.1.2, we discuss in more detail the cognitive load on the users depending on the different tasks.

Whether it is more effective to use a traditional ad-hoc retrieval system compared to a RAG system depends on the solution object for the task. We refer to a solution object (also known as solution path) as a path in a graph representing the search space with which a user wants to satisfy his information need. Specifically, a solution object can be a piece of text, a document, an image or any kind of media that satisfies a user's information need.

- If a solution object already exists for a task, such as a paragraph of a web page, a PDF document, or image, rank-based retrieval with traditional ad-hoc retrieval systems is more effective for solving a task.
- Vice versa, if the solution object does not exist, but needs to be created through the process of retrieving several information items and synthesizing the solution object out of it, then it is convenient for the user to use a RAG system. Then, the user is not exposed to the cognitive load of creating the solution object, but instead needs to specify only the request and, potentially iterate through feedback with the RAG system.

In the following sections, we resort to Bloom's (revised) taxonomy [92] to compare the cognitive tasks a traditional ad-hoc retrieval system and a RAG system takes responsibility for. Subsequently, we develop a taxonomy of cognitive tasks that potentially fall within the capabilities of RAG systems. Along Bloom's (revised) taxonomy we then discuss the possible limitations and negative effects of RAG systems on the user.

4.1.2 Bloom's Revised Taxonomy

Bloom's revised taxonomy[2], see the colored pyramid in Figure 1, provides a concrete medium to navigate through the set of tasks that a user would like to accomplish. Bloom's Taxonomy is a framework that organizes learning objectives into a hierarchy of cognitive levels, from basic remembering to complex creation. While with RAG systems tasks of all cognitive levels in Bloom's taxonomy can be addressed, rank-based retrieval addresses tasks of the two lowest levels only. This division of the pyramid also reflects our distinction between tasks for which the solution object consists of a a single information item that already exists (the two bottom levels) and for which the solution object must be synthesized (the top four levels) and tasks. In this analogy, a rank-based retrieval system has to remember the documents (search index), and understand the user's information need to be effective. On top of that, a RAG system has to apply domain-knowledge, analyze the retrieved documents, evaluate individual aspects, and create a final response to present to the user to fulfill the information need.

In the following, we present a brief overview of Bloom's taxonomy from the user's perspective. The taxonomy also provides a proxy of the required level of cognition for a task, going from lowest (*remember*) to highest (*create*).

- **Remember**: Users recall or recognize knowledge from their own memory. They rely on their memory to produce or retrieve facts, definitions, lists, or recite previously acquired information.
- **Understand**: Users construct meaning from different types of functions. These functions include written or graphic messages. Understanding involves tasks like interpreting, inferring, classifying, summarizing, explaining, or exemplifying.
- **Apply**: Users use, implement, or execute a procedure. Applying involves situations in which users rely on learned materials and information through products like models, interviews, presentations, or simulations.
- **Analyze**: Users decompose materials and concepts into components and identify (inter)relations between these components and relations to an overarching structure or purpose. Analyzing involves differentiation, organization, attribution, and the ability to distinguish between components. Analysis products include spreadsheets, surveys, diagrams, charts, and graphic representations.
- **Evaluate**: Evaluation remains an integral part of the pipeline where users make judgments based on the analysis/observations and their prior knowledge about the task.
- **Create**: For a large set of tasks, a user's final goal is to come up with a coherent/functional form of new patterns or structure, given the previous set of observations, where the primary aim remains to synthesize a new solution object. This leads to the most difficult features that a user would want a system to augment.

We consider the above-mentioned levels from the RAG users' perspective and formulate a task taxonomy for RAG Systems. Further, we also explore the challenges and limitations of RAG-based systems, which may play a crucial role in affecting users' cognitive skills at different levels in Section 4.1.4.

[2] The main differences between Bloom's Taxonomy and the revised version [92] are terminology, structure, and emphasis. Bloom's Revised Taxonomy replaces the original nouns (Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation) with verbs (Remember, Understand, Apply, Analyze, Evaluate, Create), shifts the top level to "Create", and adds a second dimension of "Knowledge" (Factual, Conceptual, Procedural, Metacognitive) this way creating a two-dimensional framework.

1. **Information Access**
   a. **Knowledge Retrieval & Organization**
      i. Question answering for complex information needs
      ii. Summarization of retrieved information items
      iii. Clustering of retrieved information items
      iv. Comparison of retrieved information items
   b. **Knowledge Acquisition**
      i. Interactive learning of new knowledge related to concepts, processes, items
      ii. Feedback and reflection
      iii. Explanation of specific concepts, facts, or events
   c. **Content Recommendation**
2. **Planning & Solution Implementation**
   a. Problem-solving (e.g., mathematics, logics, coding)
   b. Procedural instruction retrieval
   c. Planning & scheduling
3. **Content Generation & Creative Tasks**
   a. Content creation (e.g., stories, poems, images, documentation)
   b. Hypothesis generation
   c. Generation of communication items

Figure 2 Taxonomy of tasks that is preferable to be solved with RAG from the perspective of a user. All tasks in this taxonomy require the system to synthesize of a solution object.

4.1.3 Task Taxonomy

A central aspect when thinking about cognitive tasks that users want to accomplish is to distinguish between tasks for which RAG is the preferable system over rank-based search. We identified that RAG is superior for all tasks that require synthesis of a solution object, since no solution object is preexisting. In Figure 2, we provide an extensive taxonomy of all tasks potentially solvable with RAG that require synthesis from the perspective of a user.

The task taxonomy was created in a bottom-up fashion, by first collecting all thinkable cognitive tasks, that could be solved with a RAG system, and then they were clustered, classified, and organized in the taxonomy. In the following, we outline the tasks categories from the taxonomy in more detail.

4.1.3.1 Knowledge Retrieval & Organization

The primary objective of knowledge retrieval tasks is the short-term information access with well defined information needs. We define short-term tasks as tasks that can be accomplished in a single session involving only a few turns in a conversation. The central task of knowledge retrieval is open-ended question answering. With a RAG system, the system itself can fulfill tasks with more complex information needs, whereas ranked search systems would leave the cognitive load of multiple searches, comparing and synthesizing information to the user. Another core task that users expect to be satisfiable solved with RAG is summarization. This task involves either single or multi-document summarization and typical instances include key point extraction and overview preparation as well as comparison, classification, and clustering of different concepts.

4.1.3.2 Knowledge Acquisition

In contrast to knowledge retrieval tasks, knowledge acquisition tasks describe rather long-term procedures, which usually require multiple sessions to be accomplished. Knowledge acquisition is mostly concerned with supporting users to build up substantial reservoirs of knowledge. Users can learn about various topics (e.g., learning fundamentals of physics) or acquire different skills (e.g., learning how to program) in interactive learning scenarios. Different learning strategies depending on the cognitive level of the user can be offered to the user that consist of providing explanations, feedback, or encourage reflection. Retrieved information items for these tasks can include educational materials but also, for example, grading schemes.

4.1.3.3 Content Recommendation

We consider the task of (conversational) recommendation as an information access task in which the classical notion of relevance is replaced with user preference. Consequently, recommendation items can be retrieved with a search engine and ranked according to user preference. Typical recommendation tasks that users fulfill with a RAG system are uttered as a question (e.g., “what should I cook tonight?”). Thus, RAG can be seen as a cross-domain and multi-modal recommendation.

4.1.3.4 Planning & Solution Implementation

While planning tasks are to provide step-by-step plans and instructions, problem-solving tasks incorporate implementing solutions and providing the results to the user. Mathematical, logical or coding tasks are examples of problem-solving tasks that follow strict constraints. In these tasks, retrieved information items can include educational materials providing explanations for these constrains. For tasks outside of our problem-solving category, constraints can be weaker or even optional. These tasks can involve procedural tasks that can be anything from retrieving while following a cooking recipe to repairing an electronic device. Information items that can be retrieved for these tasks can be manuals, recipes, and other documents containing instructions. Furthermore, planning and scheduling tasks can be planning trips or scheduling appointments for which transport schedules or calendar item retrieval are required, respectively. For some of these tasks in this category, the RAG system also can directly implement the solution. For example the solution can be a code artifact for solving a coding problem, a sent email or a scheduled meeting if the task relates to communication or calendar planning.

4.1.3.5 Content Generation & Creative Tasks

Content generation and creation tasks relate to the top of Bloom’s Taxonomy. In the context of RAG, it includes the creation of new content, the generation of hypotheses, and the generation of communication items. For these tasks, the retrieval can be done over different types of information items, including character descriptions to write a story or generate an image, existing Lewis structure of molecules to hypothesize a new one, or messages in multiple threads to write an email. Based on the retrieved information items, the RAG system aims to generate a new information item, which is then the solution object.

4.1.3.6 Domains of Applications

Based on the pool of tasks imagined for the creation of the task taxonomy, we envision scenarios in different domains and population segments for which RAG systems can be particularly beneficial.

- For computer science engineers, with expertise in coding, a RAG system can support their daily tasks related to the creation and management of codebases. For example, the RAG system can implement new features, identify weaknesses, review code and summarize optimization strategies given the current status of a project and external documentation.
- For K12 pupils, a RAG system should be tailored to make the pupils engage with it to lead them to the solution object rather than providing it directly. For example, using different modalities when learning a new language and showing the richness of the language using paraphrasing.
- For elderly with different level of cognitive decline, a RAG system could remind them about their daily tasks, curate an image gallery of their life to help with memory and assist them in keeping their skills
- For marketing specialists, a RAG system could monitor trends for their respective market or create slides for presentations.
- For job seeker, a RAG system could give feedback on CVs, teach him/her new skills, help prepare for interviews, and monitor new open roles that come online.

4.1.4 Challenges and Limitations with respect to User's Cognitive Skills

Despite great advances and the already high effectiveness of RAG systems, the overuse of these systems imply certain risks for the user. There is growing evidence that GenAI systems contribute to the deterioration of user cognitive skills [63, 90]. Over-reliance on technology shifts many essential mental tasks from our brains to our devices leading to cognitive offloading. For example, it has been found that GPS replaces spatial navigation [183], search engines might replace memory recall [156] or AI assistants could take over reasoning and problem-solving [63]. In case of RAG systems, user cognitive skills across all levels of Bloom Taxonomy (Figure 1) might be affected as outlined below:

- **Query formulation:** Automatic query formulation replaces the formulation of keyword queries from the user's side with natural language problem descriptions. These problem descriptions neither have to be complete nor comprehensible to a great extend in order for the RAG system to provide somewhat satisfactory results.
- **Source triaging & relevance assessment:** RAG systems filter the retrieved sources before a user sees them and takes away the assessment of relevance for individual sources.
- **Close reading & extraction:** RAG systems typically provide well structured summaries that do not require deep engagement with the text to pick up key points. Generally, close reading of individual documents is discouraged.
- **Multi-source synthesis:** RAG automatically integrates retrieved evidence into automatically synthesized responses which is taken away from the users.
- **Factual recall:** RAG responses tend to be complete or at least provide the option of follow up questions that reduce the need for the user to remember individual facts or recall knowledge.
- **Critical thinking:** RAG systems typically include references to external sources which transports a sense of trustworthiness. This trustworthiness may discourage the user to

do fact checking and external validation, which leads to a decline of critical reasoning abilities.

- **Planning & creation:** RAG systems can provide step-by-step instruction on how to solve a problem. Thus, users do not need to develop their own solution strategies or a novel solution object, which might hinder creativity, or development of planning and problem-solving abilities of an individual.

We expect that the push towards efficiency and the increased ease of use of LLM and RAG systems will result in a reshaping of the cognitive landscape of users, and likely in less involvement of users' cognitive processing in reaching the output, as well as in various forms of superficial learning and shortcuts. While potentially more information can be provided by RAG systems and consumed by users due to the systems' higher coverage of retrieval sources compared to what a user can find during her search as well as intrinsic and extraneous cognitive load [157] of reading can be decreased, the information is usually already extracted, synthesized, processed, structured, and presented for the user. Furthermore, GenAI poses a risk of standardizing language and reasoning for its users [155], potentially presenting a limiting factor for diversity and richness. While in Section 4.1.3.6, we name examples of ideal cases where RAG systems could help users, the risks should not be ignored, especially for vulnerable groups like children and the elderly, who could be adversely affected when it comes to their cognitive development and engagement.

We argue that several scaffolding elements could be incorporated into the design of future RAG systems to make them support user cognitive engagement in relevant settings, such as user learning and training. While these alterations would typically lead to diminished efficiency and effectiveness in reaching direct outcomes/answers, this would come with an added effect of fostering cognitive abilities of users and a potential increase in trust of the RAG system (e.g., by "revealing" the solution process). Such scaffolding elements could be then especially suited for cases when speed, ease of use, and accuracy may not be the most important, or could be sacrificed, much like jogging and physical training result in an increased effort and higher strain on those who prefer these activities rather than moving by cars.

We highlight below a set of such scaffolding elements that align with each of the identified cognitive skills affected.

- **Make retrieval more visible and interactive to keep information discovery skills.** For example, the system can show queries used by RAG and perhaps even allow the user to add/modify them, thus exposing the retrieval stage. The system could then explain to users how documents were found and chosen, or let them intervene in this process.
- **Promote close reading and evidence engagement to preserve critical reading, comparison, and fact-checking skills.** The system could highlight where particular information nuggets come from, quote some parts from original sources, or direct users to read selected content verbatim. In addition, the system could include a reflection stage or could prompt the user to think critically.
- **Let users integrate multiple sources to support information synthesis skills.** The system could visualize agreements and conflicts across sources, or suggest alternative summaries and different synthesis paths for comparison.
- **Embed counterfactual reasoning to support reasoning and critical thinking.** The system could provide one-click generation of alternative perspectives or assumptions, and/or show reasoning chains to let the user edit assumptions or steps. Moreover, the

system could provide an option to change an assumption and see how the conclusions differ in result.

- **Gamification, hinting and user questioning.** Instead of providing a direct final answer, the system could add hints, incomplete reasoning paths, or quizzes. In addition, it could incorporate what-if branches to modify the text.

While many of the above-proposed adaptations have not been developed and applied yet, there are already several promising initiatives worth noting. For example, a new feature called *Study and Learn*[3] has been recently added to ChatGPT. In the above mode, the LLM lets users actively learn through guiding questions and prompting reflection instead of the provision of direct answers. For instance, a student working on a given task, can be suggested to attempt a solution first or to explain reasoning before the answer is revealed. Kazemitabaaret et al. [83] propose CodeAid, a system that answers conceptual questions, generates pseudo-code with line-by-line explanations, and annotates student's incorrect code with fix suggestions instead of using directly the LLM to provide a coding solution. Jangra et al. [75] advocate the use of hints for letting users find the answers to their questions by themselves with the help of LLM in the form of generated hints. There are already several datasets of hints available, both manually created [119, 175] or automatically generated [76, 120] along with their corresponding questions and quality metrics. Toolkits like HintEval [121] provide methods for hint generation and evaluation measures such as convergence which estimates the ratio of candidate answers to the user's question that can be discarded after the application of a given hint. In [128], the authors propose a hint generation framework for middle-school level math word tasks using a language model to decompose the solutions into atomic mathematical operations.

The effect of any design improvements on the cognitive adaptation and intellectual effort of users need to be however continuously and carefully studied [182] based on longitudinal user studies, or even using electroencephalography (EEG) signals [90, 99]. Additionally, the benefits coming from such adaptations should be compared to the reductions in task completion efficiency and effectiveness. It is also crucial to differentiate the analyses based on user attributes such as educational background, age, culture or on learning objectives, as well as on the cognitive demands tasks inherently present.

4.1.5 Research Questions

Despite the progress in LLMs and RAG systems, many aspects of RAG still need to be improved to fully serve users as desired with their cognitive tasks. Below are the research questions that arise to meet the desired RAG portrayed in the previous sections.

1. **Understanding the user** RAG relies on LLM's capability of natural language understanding to understand the user's question and determine what information should be retrieved. In some cases, the question may be wrongly understood or the retrieval query wrongly formulated. A more accurate understanding of the user's question/intent and better formulation of search queries can improve the quality of RAG answers. In addition to improving LLM's general capability of language understanding, RAG-specific research questions can be: How to improve query formulation? How to leverage context information in conversation when RAG is embedded in a conversational system? How to involve the user in the loop to better formulate the question/query?

[3] `https://openai.com/index/chatgpt-study-mode/`

2. **Find relevant information when needed** An important task in RAG is to retrieve the most relevant information to support answer generation. Currently, the retrieval module is generally loosely connected to the generation module. The retrieval module is often optimized separately using traditional ranking objectives. An interesting research question to investigate is whether different retrieval algorithms or systems should be developed so that the retrieved information fits better the LLM for generation, instead of fitting human users. This problem is further discussed in Section 4.3.2. The retrieval algorithm can also be optimized for answer generation, i.e., a direct feedback from the generator can be used for the optimization of the retriever.
   Another interesting question is to detect the need for retrieval. Many questions can be reliably answered by LLMs, so retrieval is not always required. Retrieval is needed only when LLMs may lack sufficient knowledge on the question, or need to retrieve additional evidence to confirm an answer. Research in this line has started [13, 79, 159], but the detection needs to be more accurate.
3. **Measuring the quality of information** IR algorithms have been optimized to do ranking. There is no or little indication on the quality of the information retrieved. Questions that arise may be: Is the information relevant? Is it from a reliable source? Is it trustworthy or authoritative? Is it worth being integrated with LLM for answer generation? These questions have not been extensively investigated in IR literature. There is a need to assess the quality of the retrieved information for LLM generation.
4. **Integration of retrieved information in generation** The retrieved information is commonly added as part of the context when asking LLM to generate an answer. This loose coupling mode leaves full freedom to LLMs to use the information in its own way. Stronger coupling involves more interactions between the two modules 4.3.2. Despite the fact that LLMs can be finetuned to better integrate the information, there is no guarantee that the integration is well done. The integration is straightforward when the retrieved information is consistent with or complementary to the internal knowledge of LLMs. However, how about when it is inconsistent/contradictory with the internal knowledge? when the information is irrelevant? when it is from a unreliable source? This question is also discussed in Section 4.2.6.
5. **Contextual factors** RAG will be used in different application contexts (for general QA, or in education) and for different users (children, adults, students, ...). The information to be retrieved and the type of answer to be generated should be adapted to the context and the user.
6. **Reasoning capability** Many of the questions cannot be directly answered by retrieved information. It is necessary to connect or compare the information or perform reasoning. Although the capability of reasoning of LLMs has much increased in recent years, it is still insufficient for answering complex questions or achieving complex tasks. Developing higher capability of reasoning based on the retrieved information is still a pressing research topic.
7. **Specialized knowledge and skills** As RAG is expected to be applied in a variety of application contexts (coding, mathematics, medical applications, ...), it is necessary that it exploits domain knowledge and expertise, which can be presented in various forms: as texts, knowledge graphs, specialized LLMs, or a set of tools. Despite recent progress in these areas [8, 109, 192], more research is needed to find better ways to leverage them in RAG.
8. **Knowing its limit** However a RAG system is powerful, there will be always questions that it cannot answer, either because there is no answer or it does not know the answer.

Admitting “I don’t have the answer" requires RAG to be aware of its own limits. More research in this line will also reduce the likelihood of hallucination.

9. **Configurable RAG systems** In many special applications such as in education, a RAG system should be configured to retain its final answer, but provide a form of partial answer or hint to help the user to think. This question has not yet dealt with in research.

4.2 Information Behavior and Result Presentation in RAG

Mohammad Aliannejadi (University of Amsterdam, NL)
Gianluca Demartini (The University of Queensland – Brisbane, AU)
Carsten Eickhoff (Universität Tübingen, DE)
Norbert Fuhr (Universität Duisburg-Essen, DE)
Sebastian Heineking (Universität Leipzig, DE)
Martin Potthast (Universität Kassel, DE)
Harrisen Scells (Universität Tübingen, DE)
Damiano Spina (RMIT University – Melbourne, AU)
Johanne Trippas (RMIT University – Melbourne, AU)

The working group on information behavior and result presentation focused on user diversity, task variation, and interaction modalities. RAG systems benefit from natural language interfaces and grounding outputs in external knowledge, enhancing transparency and efficiency. However, challenges remain around over-reliance, critical thinking erosion, and cultural bias from dominant sources. We outlined the need for adaptive architectures that support diverse users—ranging from experts to vulnerable populations—and varied tasks, from factual queries to creative generation. Three key dimensions of interaction were outlined: user/system initiative, information progression and complexity, and human-machine collaboration – highlighting gaps and opportunities in current systems. Four collaboration paradigms were proposed, emphasizing future potential in collective intelligence. Open research questions aim at designing adaptive, transparent, and inclusive RAG systems that balance automation with user agency and engagement.

4.2.1 Motivation

Before the advent of the Web, libraries and academic institutions relied on computerized bibliographic databases accessible via terminals, including citation indexes, abstract journals, and specialized database subscription services. Users searched by subject headings, keywords, or author names, often helped by librarians [53]. The result of these searches was a list of references, followed by the necessary steps to access the actual documents. With the advent of the Web, the compilation of such reference lists has been automated, and with the subsequent rise of digital libraries, the documents themselves became directly accessible. Web search engines have become the gateway (and the gatekeepers) to information on the Web. The primary mode of interaction with them is a short keyword query or question as input, and a ranked list of documents as output, from which searchers can extract the desired information. Following many recent breakthroughs in the training of large language models, retrieval-augmented generation (RAG) systems are now a new form of information system that, based on using a traditional search engine, not only retrieves relevant documents but also reads, extracts, and aggregates the information sought in the form of long-form text [144]. As a result, a significant portion of the cognitive effort associated with information retrieval is shifting from the user to the information system [74]. Consequently, a fundamental change in the information behavior of searchers is to be expected.

Industry develops and deploys prototype AI-based systems in a manner of rapid prototyping at scale today, the initial mode of interaction with was based on the basic chat interface—a natural way to interact with language models. This interface has prevailed so far and has become a current de factor standard for interactions with RAG systems and search result presentation. So far, the new search interface can be characterized as follows: an input mask for long, instructional, and complex question- or task-related queries; an output in chronological order of successive queries; a contextual dependency of new responses on previous ones; and the option to delete the context and start a new chat. This interface goes beyond the classic search engine results page by explicitly taking contextual information into account in a way that is largely transparent to the user. Session-based search is thus the norm in chat-based search interfaces, whereas in the traditional ad hoc search interface, session-based search has played an implicit role at best, but usually none at all.

In transition from a traditional search engine to a RAG system, the information behavior of users changes [168]. In what follows, we briefly review query input, the use of search operators in queries, and the consumption of search results.

When using RAG systems, queries become longer, more complex and increasingly specific. Unlike traditional search engines, RAG systems much better understand the user's information need or their tasks. Moreover, users no longer need to learn a query language (i.e., to translate their request to a set of keywords, sometimes combined with search operators). Nevertheless, certain prompting techniques may establish themselves as a new form of query language. RAG systems moreover support multimodal input, especially voice input, which makes them accessible to a broader range of users, or in contexts where the interaction with a search engine has been cumbersome (e.g., while on the go). In addition, users are increasingly describing the context in which their information need arose as part of their queries, greatly improving query understanding [135]. This can go so far that only the context, such as the goal to be achieved or the problem to be solved, is specified, while the actual information need remains implicit.

The use of search operators has declined to virtual non-existence for the majority of users in traditional search over the past decades [94]. Although they still can be used, they have gradually been hidden in sub-menus and thus fallen out of sight for most users. Most users are not aware that operators such as phrase search or Boolean connectives exist. In queries to RAG systems, there are no pre-defined operators whatsoever. However, users can invent and specify basic or even complex operators ad hoc, for example by asking the RAG system that all generated statements be substantiated with direct quotes from the sources, that the output be in a specific format, or that the results fulfill specific properties. However, since such operators are not standardized nor recommended to the user, the majority will not formulate any when using a RAG system. At best, users will exchange effective search strategies via word of mouth.

User behavior when consuming search results is changing fundamentally on RAG responses (text SERPs) compared to traditional search engine results pages (list SERPs): On traditional search engines, users skim the search results to select documents to visit [66]. This is aided by a standardized design of each search result, namely the visual highlighting of different parts of it, showing a query-biased snippet extracted from the document and highlighting the query terms in them [73]. The long texts that result from RAG systems are now either read linearly with full attention or, more likely, also skimmed to identify the specific passages of the generated text that address the user's knowledge gap. Targeted navigation support through query-biased text highlighting, analogous to keyword highlighting in snippets, is currently missing in RAG responses.

In longer search sessions, users converse with the RAG system and collaboratively "develop" an answer to their information need or a solution to their task. This information behavior has no precedent in traditional ranked retrieval systems.

The search results underlying the generated answers are displayed much less prominently than before. Their presentation follows the citation style of scientific texts, resembling number or sometimes name citations. This may include also interactive features, like displaying details similar to a traditional search result by hovering with the mouse over them. Alternatively, the search results are listed (in collapsed form) below the text. In the case of name citations, which often name the website's domain, the user may get an idea of the source and can potentially assess its reliability or authority based on their own experience. Whereas traditionally one or more search results were clicked on and the linked pages visited—unless the snippet already satisfied the information need—the new display reduces the user's motivation to explore (cf. studies on the click behavior of Wikipedia readers on citations [117]).

The information obtained from the sources to generate a RAG response is not quoted verbatim, but rather adapted to the context or the flow of argumentation of the generated result and often rephrased in an abstractive manner. Unlike extractive snippets on traditional search result pages, the statements preceding a citation cannot be found directly in the cited source document, which makes it difficult to verify the claims contained in them. In the long term, this could discourage even users who are willing to verify statements. In general, LLMs and RAG systems do not quote verbatim on their own initiative, apart from reproducing training data. While early RAG systems were not technically capable of doing so, today's systems can at least be explicitly instructed to include verbatim quotations, which is a potential search operator.

If a satisfying answer or solution to the user's need or problem cannot be found, the conversation is terminated and, if necessary, restarted without context. If a successful solution is found, a conversation can be terminated prematurely as well. It remains unclear to outsiders how the two situations can be distinguished, for example by analyzing user behavior through observation of scrolling interactions. Many models ask after each turn whether to proceed with a relevant next step or whether the user is satisfied. Since some users use polite forms such as "please" and "thank you" in their queries, transferring social conventions of human interaction to communication with the machine, it is plausible that some of the answers to such questions are short, explicit relevance feedback such as "Yes, that's not enough for me" or "No, I'm satisfied." However, in the absence of query logs, this remains speculative.

In many ways, interacting with RAG search systems is similar to giving a research assignment to a reference librarian in a library or to a personal assistant who provides concrete answers, possible courses of action, and relevant evidence. Even though current search interfaces significantly expand the possibilities for searchers compared to conventional search engines, there is still great potential for developing interfaces that better support users. Below, we provide an overview of the results of a systematic exploration of the possibilities offered by LLMs and other generative AIs.

4.2.2 Users and Tasks

A constraint of RAG system is that the input is provided by the user (and possibly their context). Thus, the RAG system is responsible to provide the appropriate output given the input. One challenge with this is that users may have different types of tasks and information needs in mind (see Section 4.1 for more details). For example, a RAG system may be used for factuality or for creativity tasks and different search and generation strategies may be more appropriate for different user objectives.

It is also important to note that different users may need different types of RAG systems and interactions modes. For example, young users, elderly users, vulnerable groups, domain experts, low-resource users may all benefit from different types of RAG system architecture and interaction frameworks.

4.2.2.1 Benefits

The naturalness of the interaction mode (input and output in natural language) of existing RAG systems has made the entry bar substantially lower for a number of users. Delegating the task of summarizing search results to the machine, it may be seen as a way to access and consume information more *efficiently* (as some of the cognitive effort is offloaded to the system).

Compared to non-RAG conversational assistants, one of the main advantages of RAG is that provides a mechanism to ground responses to external knowledge, thereby reducing the risk of providing incorrect – or outdated – information generated using the internal knowledge of the LLM. As the response is generated by combining both external and internal knowledge, the system can attribute elements in the response to sources, providing a more transparent way to show the provenance and the reliability of the information. Related studies have shown how showing the plan an AI agent intends to follow can increase trust in the system [69].

4.2.2.2 Risks

In most cases, users want to minimize effort and maximize gain. This comes with the risk of over-reliance on the RAG system and of accepting the answers given as good ones. This may, in the long term, reduce the overall ability of users for critical thinking. A possible approach to deal with this is to design RAG systems that keep their users alert, vigilant, and critical of the output received from the RAG system. This could be achieved in different ways, e.g., with systems better communicating uncertainty, quality of the sources, and proposing alternative answers (see Section 4.3.3).

Another risk is that of using dominant sources of information for RAG. This may lead to cultural bias and lack of inclusivity of diverse points of views (see Section 4.4.7). To deal with this, future RAG system may need to consider a diverse set of sources to make sure different viewpoints are included in the generated response.

For different type of tasks (see Section 4.1) different levels of reliance on sources may be appropriate. For example, for information access tasks (e.g., learning) a strong dependence on reliable sources of information is necessary while for content generation and creative tasks (e.g., hypothesis generation) a higher degree of generated content may be appropriate.

4.2.2.3 Opportunities

As the population of users grows, it is also increasing the diversity of needs and tasks. It is important to consider how we can characterize such diversity, to be able to conceptualize and design RAG solutions that accommodate for such needs. This has implications in the tailoring of all different aspects of RAG user's experience: customization, ease-of-use, functional fixedness, cultures of interaction and reasoning, etc. For instance, web search engines provide an *Advanced Search* option that provides a substantially more powerful way to control the behavior of the search engines; however, this advanced functionality is rarely used by most of searchers. If we design more complex interfaces for RAG, which users or tasks will these interfaces support?

Similarly to other information access scenarios, users of RAG systems still need to go through different stages when forming their information needs [165]. There is an opportunity to understand how interactive RAG systems can better support users during the realization of their information needs – not only for information access tasks, but also for more complex tasks such as planning and content generation tasks.

If we consider that we will move beyond the "one-fits-all" paradigm, this has also implications in the way we should conduct our research. Participatory research initiatives will be needed to be able to gather requirements and characterize needs of different groups of users. We will also need to account for richer test collections and novel experimental methodologies to better understand which UI or mode of interaction is more effective for (or preferred by) which users.

4.2.3 Open Research Questions

- How do we design RAG architectures that foster users' critical thinking abilities?
- How do we design RAG architectures that keep track of the provenance and lineage of information to transparently surface it to the end users?
- How should RAG system adapt to different users and tasks? This includes users who are domain experts and have complex information needs related to their domain, younger or elderly users with safety requirements, and users with diverse information access tasks.
- Should RAG system result presentation be adaptive and dynamically adjust for different users and information needs (e.g., learning and fact finding vs creative tasks) or should we rather have different RAG systems for different purposes?

4.2.4 Interaction Modes

RAG presents a wide range of possibilities to interact with information access systems. We believe that the mode of information with a RAG system heavily depends on three aspects: (1) the roles and initiatives of the user and the system, (2) the progression of information and the density of information at each information processing stage in the RAG system, and (3) the amount of collaboration between users and RAG systems. These three aspects define how information should be used and presented within a RAG system, and their combination can be used to define the tasks that are possible. We start by identifying the benefits and limitations and problems that exist in typical RAG systems of today before elaborating on the above three aspects, and finally identifying the opportunities that exist regarding modes of interaction with RAG systems.

4.2.4.1 Benefits

Existing RAG systems allow for natural language as the main form of interaction. This form of interaction is highly expressive and most interfaces are built around this mode of interaction. Furthermore, chat-like interfaces are already familiar to most users. Drawing parallels to list SERPs from search engines, text SERPs and chat-like interfaces provide a linear interface to interact with information, which further reduces the burden on users.

4.2.4.2 Limitations and Problems

However, this one-size-fits-all approach to interaction may limit how information is accessed by users, and other modes of interaction may be more suitable depending on the context and task at hand. For example, pointing at or marking up information may be a more natural way of selecting to learn more about or combine information.

4.2.4.3 Roles and Initiatives

The first dimension defines the level of initiative by the RAG system and its role in the interaction with humans. This dimension combines the concepts of (i) intervention in information behavior [36] and (ii) system involvement [19].

Coghlan et al. (2025) [36] define four models of search engines, ranging from a low to a high degree of intervention. The *Customer Servant* gives the user of the search engine *exactly* what they asked for without further clarification or value judgment. In contrast to that, the *Librarian* tries to discern what the user is *actually* looking for by consulting additional information like past search behavior or asking clarifying questions. The *Journalist* tries not only to understand the actual information need, but supports the user in gaining a more wholistic understanding of the topic underlying the information need. As examples, the authors cite that a *Journalist* would outline multiple perspectives and arguments on a topic or provide debunking information in response to conspiracy theories. The *Teacher*, as the highest level of intervention, focuses on the learning experience and tries to encourage critical thinking. As a consequence, a *Teacher* will not rank results only based on their relevance to an information need, but also on the basis of their deemed correctness and educational value.

While the degree of intervention is primarily concerned with the information behavior of the user, system involvement focuses more strongly *who* defines and executes search commands: the user or the system. Bates (1990) [19] suggested four levels of system involvement in search as well as a level 0 with no system involvement. At the first level, the system offers passive support. It acts as a knowledge source to provide information when prompted, but does not monitor user behavior or execute a search on its own. Examples of information include explanations of commands or suggestions for search strategies. Systems on the second level execute both individual queries as well as full search strategies when prompted by the user. The system is still not proactive but reduces the procedural load on the user by executing more complex operations like a search in all publications of a journal for a given period. The first proactive level is level 3. Here, the system monitors the search behavior to offer suggestions even when not prompted to do so. This can range from simple suggestions like removing typos to complex ones like trying to infer the underlying information need from individual queries and suggesting a search strategy based on that understanding. On the highest level of system involvement, level 4, the system conducts search activities on its own. Based on user preference, the system can report its performed actions or only present the results.

Bates defined the level of system involvement as the complement of user involvement. In other words, as the involvement of the system increases, that of the user decreases. This zero-sum setting in involvement is the reason why we propose to couple it with the degree of intervention in information behavior. A highly involved RAG system bears the risk of users delegating large parts of their cognitive work to the system as it automatically executes tasks. With a higher degree of intervention, a RAG system might be able to keep the user involved and decrease the risk of cognitive offloading.

Figure 3 gives an illustration of the level of initiative that a RAG-system can take when interacting with a user. It is important to note that the same system can exhibit different levels in different situations. On the left side of the spectrum, the system executes searches only when prompted by the user and does not intervene to clarify the information need or present multiple perspectives on a topic. At a higher level of initiative, the system does not simply execute a query or command, but asks questions *after* completion. While the system still only acts upon explicit request by the user, it may list suggestions for follow-up questions (level 3 of system involvement) and clarify if the presented information meets the

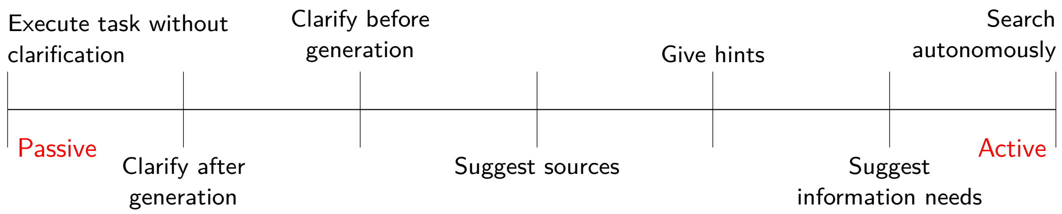

Execute task without clarification
Clarify before generation
Give hints
Search autonomously
Passive
Clarify after generation
Suggest sources
Suggest information needs
Active

Figure 3 The dimension of initiative taken by the RAG system with examples for different levels.

requirements of the user (like the *Librarian*). This is the level of initiative taken by most of the current RAG systems.

Instead of trying to complete the user request based on the most probable interpretation and clarifying *after* completion, a system of higher initiative performs the clarification step *before* taking any actions. As one advantage, this allows systems to handle cases of uncertainty about the information need, giving the user the opportunity for clarification instead of requiring them to verify retrieved content or generated output that was based on a false assumption. In addition to clarifying the information need, the system can also ask for confirmation about a planned course of action like the sources to be used or the sequence of steps the system will take to perform a task. While this form of initiative is still on level 3 of system involvement (no proactive search), the degree of intervention can reach that of a *Teacher* search engine, for example when the RAG system asks the user to consider a different course of action. As a consequence, this level of initiative needs to be employed carefully as to not make the system appear patronizing or annoying because it "refuses" to execute the user's request. The same is true for all following levels of initiative as the system becomes both more involved in the search process as well as intervenes more strongly in the user's information behavior.

A next higher form of initiative is for the system to suggest new sources to the user. Based on the history of requests, the system may identify useful sources to be added for future requests. While previous examples are limited to initiations by the user, i.e. they occur in response to a query or command, the source suggestion can be executed without the user asking for it.

Giving hints is a level of initiative that falls completely into the domain of the *Teacher*. The idea is to guide users toward answers without directly revealing them, encouraging deeper engagement with a topic and trying to mitigate risks like information overload and overreliance on automated reasoning [121]. A hint-based interaction can be either started by the user prompting the system or initiated by the system based on prior information behavior. Hence, it can be located on level 3 or level 4 of system involvement.

An example of high initiative is for the system to not only recommend sources, but information needs. As the user interacts with the system over time, a growing history of information needs and prior requests is collected. In contrast to traditional search engines, the user does not only give implicit feedback in the form of clicks, but explicit feedback in natural language. This allows the RAG system to analyze and relate previous information needs in great detail to make predictions about new information needs that the user has not yet articulated. This happens without prompting by the user.

At the highest level of initiative, the system performs searches autonomously. While the system involvement is by definition that of level 4, the degree of intervention can be both that of a *Journalist* as well as a *Teacher*. In the former case, the system would retrieve results from

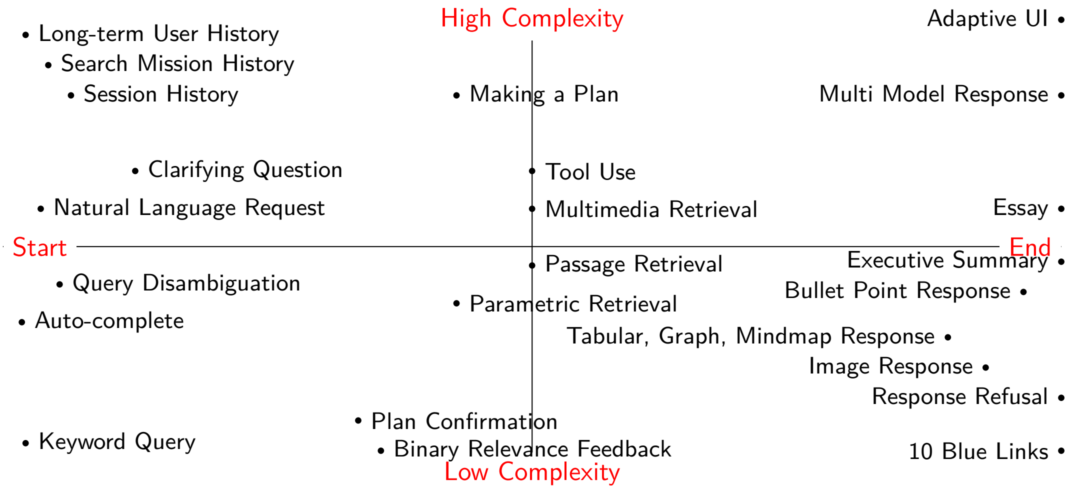


Figure 4 The two dimensions of information progress and information density, mapping some of the possible types of information that are used and produced by RAG systems.

various perspectives to present a topic as wholistic as possible. The *Teacher* would go beyond that and not only present relevant results, but curate them based on educational value. In that role, the system would also ask questions or reveal new information step-by-step to teach the user about the topic.

4.2.4.4 Information Progression and Complexity

The second dimension is that of information progress and information complexity, which are tightly coupled. From these two dimensions naturally arise the types of information that are used and produced in a RAG system. Figure 4 illustrates this relationship between these two dimensions and some of the possible types of information. We define information progression as the stages of the information processing, for example, the start of interaction of a RAG system could be a user submitted query, while at the other end of the information progression spectrum could be a generated response from the system. We define information complexity as the amount and intricacy of information being processed at a particular stage in the progression of information, for example, low complexity information includes a keyword queries (à la Web search) and traditional 'ten blue links' search engine results pages; meanwhile high complexity information includes the past history of interactions with the RAG system (see Section 4.2.5) and a multi-modal generated response. These two dimensions are tightly coupled with each other, and we use this coupling to identify gaps and deficiencies in existing RAG systems that restrict access to information.

First, we use these two dimensions to define gaps in interaction modes with RAG systems. For example, in Figure 4, the vast majority of RAG systems allow only for relatively low forms of information complexity as the mode of initiating interaction with a RAG system and provide only relatively low forms of information complexity as information that can be interacted with as output; the system receives a user request in (mostly) text form and the system provides a (mostly) textual response. One clear gap that we see is to incorporate more feedback from the user (depending on the role and initiative of the user). Between the input and output, on the input side, systems could actively request more information

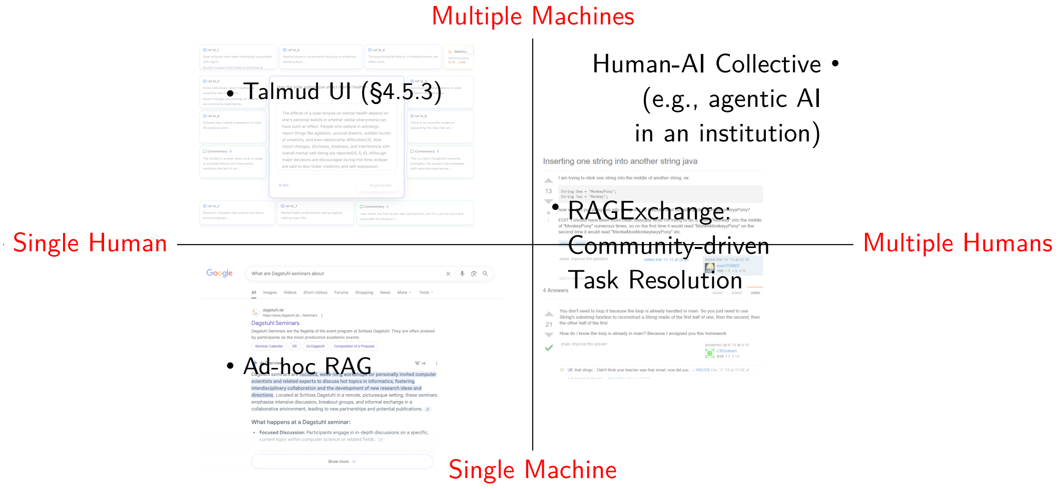


Figure 5 Human-machine collaboration paradigms in RAG.

or clarify the request. Towards the output side, systems could actively present a plan and ask for feedback on its execution. On both the left and right hand side of the progression spectrum, this feedback can either be low complexity (clicking thumbs up/thumbs down), or high complexity (written feedback to the model).

Secondly, we use these two dimensions to critique existing RAG systems, and highlight deficiencies of these systems that would allow users to better support their information access tasks. One clear critique is the relatively low complexity on the extremes of the information progression despite the relatively high complexity of information used (e.g., complex tool use). There is an opportunity to better support users in their information access tasks by allowing them to provide lower complexity information and having the system produce high complexity information. One example of a hypothetical system that addresses this critique could be one that processes simple keyword queries into entirely new user interfaces, e.g., generating interactable Web pages to support information access tasks (à la Wikipedia articles) or generating an editor interface to support creative tasks like hypothesis generation.

4.2.4.5 Human-Machine Collaboration

The third dimension of interaction modes is that of collaboration between human(s) and machine(s). We structure the following part under four paradigms of human-machine collaboration (or hybrid intelligence [4]) in RAG (Figure 5).

Single human, single machine. This is the most common paradigm nowadays in ad-hoc RAG, where a user interacts with a system to satisfy their information need.

Single human, multiple machines. In this paradigm, a user is able to work with a number of tools that cooperate to either solve the task or critically analyze how other systems aimed to address the task. Instances of this paradigm may include federated RAG (Section 4.3.5) or the Talmud UI (Section 4.5.3).

Multiple humans, single machine. While a RAG system may enable users to tackle more complex tasks, users may still struggle on how to use the system. Community Question Answering platforms are effective in collecting and sharing knowledge across users with a common goal. One may think of a RAG system (e.g., “RAGExchange”) where multiple users work together to collectively learn or solve a complex task by interacting (either synchronously or asynchronously) with a common system.

Multiple humans, multiple machines. The paradigm above is not necessarily limited to one single machine, and it is intuitive to think of multiple users using multiple systems to solve a common task collectively (e.g., tackling creative tasks, such as using multiple RAG systems to generate new content). The highest level of collaboration is reached when we have fully connected network of humans and machines, where machines also collaborate among themselves. For instance, one may envisage a human-AI collective, e.g., members of an organization or institution working together with multiple RAG systems, agents, and tools.

These paradigms illustrate the spectrum of collaboration in RAG, from individual use to rich human-machine collectives, with some already partially explored in existing systems, and other pointing to plausible emerging collaborative scenarios.

4.2.4.6 Research Questions

Interplay Between User and RAG System Initiative Both RAG systems and users can have different levels of initiative at a micro and macro level depending on the task, information progression, information complexity, and level of collaboration. RAG systems should be adaptable to these different modes of interaction, and the perceived initiative of the user.

How Information Progresses Through a RAG System RAG systems are able to make use of information at varying levels of complexity throughout the entire information processing pipeline. However, there are many forms of information that a RAG system can use, and it is not clear how long chains of information processing at varying levels of complexity impact users. There is a clear knowledge gap regarding how RAG systems use information of varying complexity as it flows through the system, but also regarding the understanding of the interactions between the different forms of information.

Sequential versus Parallel Information Processing Naive RAG systems use information in a sequential manner: the user request is used to search for information and generate a response. However, depending on factors such as task complexity and level of initiative, the RAG system could attempt to process information in multiple ways, all in parallel: one component could be responsible for proactively searching a document collection, while another component clarifies the task and yet another component works on making a plan of action. Each of these parallel information processing steps can be used to better understand the task and better assist the user in achieving their information access task.

Collaborative Modes of Interaction in RAG Systems The most common form of interaction with a RAG system nowadays is single human, single machine. However, the other combinations of humans and RAG systems allow for much richer forms of interaction with these systems, potentially offering the ability to enable users to better achieve their information access tasks. While we have presented some examples for what these modes of interaction might look like, we envision that the aspects of interaction modes discussed in this section have a role to play in the development of systems and interfaces to support highly collaborative RAG systems.

4.2.5 Keeping Track of History

4.2.5.1 Motivation & Benefits

As search tasks persist over time, information needs recur, or audit trails are required for compliance, maintaining explicit histories of previous inputs, interactions and outputs becomes desirable. Concrete benefits of such functionality are the ability to (1) re-find previously encountered material, (2) continue tasks extending across multiple disconnected sessions, (3) interrogate explicit logs of previous interactions, (4) maintain lifelogging records, or (5) try to understand the inner workings of the system.

4.2.5.2 State of the Art

There currently are no dedicated user interfaces for RAG-based history management. Most commercial language modeling services available on the market offer recency-sorted side-bars in which a linear list allows navigation into the raw historic interactions between user and system. The session contexts can then be returned to, in order to continue longitudinal tasks or retrieve information from the discourse. Several platforms offer privacy modes under which logging or platform-wide use of the interaction are temporarily suspended.

4.2.5.3 Risks

Maintaining any form of user- or system-accessible history comes with a number of potential pitfalls. In the following, we discuss four risks: threats to user privacy, failure modes in service personalization, service deterioration due to context noise, and potentially lacking user acceptance.

The most obvious risk is the potential for abuse and linking of sensitive private information. With growing duration and richness of the records maintained, the potential for harm (e.g., in case of hacks, data breaches or simply inadequate access control) to the user's privacy grows. Considering a scenario in which advanced system capabilities lead to regular reliance on the tool, its historical record would form a near-complete *virtual twin* of the user.

Going beyond privacy, a rich record of historical interactions is a valuable resource for personalization (see discussion below). As with all adaptively learning systems, such an evolution of system behavior in response to implicit usage patterns introduces the risk of unintended feedback loops, forming self-reinforcing echo chambers of biased information. It is even conceivable that potentially addictive properties may emerge over time.

Next, if historical information is used as additional context to future system interactions, the challenge arises of selecting the relevant session-orthogonal "slices" of history to be used for contextualization. Failure to adequately perform this non-trivial sub selection may lead to rapid deterioration of service quality in response to context noise.

Finally, if advanced user interfaces for history keeping (e.g., along the lines of the outlook below) are introduced, we will face challenges of equitably ensuring accessibility of tools and visual metaphors among disparate user groups. This creates a risk of providing functionality only for users of select levels of technology literacy or education.

4.2.5.4 Research Questions

In keeping with the previously discussed risks, we see a wealth of opportunities with considerable potential for beneficial innovation beyond the state of the art. in the following, we motivate nine research questions centering around the notion of history in RAG systems.

Aggregation and Summarization Long system-generated outputs and multi-turn interactions can quickly become overwhelming for the user. Returning to the extensive raw history, a long-suspended task may in fact become daunting. Instead, aggregates and summaries of interactions and outputs may offer a much smoother re-entry. Such summaries do not necessarily have to be limited to the textual modality but might benefit from visual depictions of the concepts covered and materials visited (or retrieved but skipped) during prior sessions. Finding the right (mixture of) modalities for summarization will depend on properties of user and task.

Mapping Information Landscapes As an extension to the previous research question, we encourage an exploration of explicit charts of the subspace within the collection that has thus far been explored. This could include visualizations of document linkage graphs, topical ontologies or other relevant structures induced by the concrete collection indexed for retrieval. Offering users a rich understanding of the topology and connectivity of the collection may offer considerable benefits in usability and allow for easier scrutinization of generative system outputs.

Resource Management The availability of such rich, potentially interactive aggregations and visualizations of the collection, resources, and tools available to the RAG system offers the unique potential for the user to take a more active role in resource management. This could, for example, follow the create, read, update, and delete (CRUD) [113] paradigm, allowing users to save retrieved resources for future retrieval, marking them up as "high value", creating new content in their personalized local collection, or deleting encountered material that they deem of low utility or even harmful. This final step can extend to the removal of entire information channels (such as, e.g., email in a particularly noisy inbox). It remains an open challenge how to offer this functionality without cluttering the interface and with the necessary safeguards for users to conveniently undo or modify previous actions.

Handling Recurring Information Needs A non-trivial proportion of searches are dedicated to recurring information needs. Advanced RAG system histories should include functionality for automating recurrence or re-finding of previously encountered material and answers. It is an open challenge to find effective interfaces in which to manage such needs. There might for example be an explicit register of previous interactions with external retrieved material that should be made available to both the user as well as the system, when prospectively formulating or answering new needs.

Living SERPs As an extension to recurring information needs, there are standing queries for which the user seeks periodic updates from a changing collection. Currently, such standing needs are addressed only by fringe tools. An explicit history feature in the form of a "living document" might be much more effective at serving this purpose. We can, for example, imagine self-updating documents that reflect the latest state of material available in the collection, as new scientific papers are published and existing ones might be revised and altogether retracted. The exact interface, highlighting and notification modes for such living SERPs remain an exciting challenge for future research.

Personalization From a system perspective, explicit histories offer the potential for tailoring outputs and interactions to the specific preferences and needs of (groups of) users. Importantly, this can include, but should not stop at, the level of traditional search personalization for better document relevance estimation. Instead, we can imagine sophisticated agentic workflows in which tailored interaction schemes or user interfaces are generated to fit the concrete user. We believe that it is important to communicate these actions to the user and give them control over the applied forms of personalization. This might, for instance, be achieved via additional (parallelized) retrieval operations from instruction history to arrive at preferred output formats.

Transparency and Explainability To ensure highest fidelity and trustworthiness of RAG-based systems, their histories must afford users a transparent view into the tools, resources, and operations used to satisfy their request. While there have been significant advances in mechanistic interpretation of generative systems [10, 30, 106], communicating their insights in a form accessible to users of different technology literacy levels remains an important open challenge.

Reasoning Traces A specialized form of explanation is concerned with the concrete reasoning steps and tool call that a model takes to arrive at an answer. These so-called reasoning traces are, if at all available, often returned in the form of intermediate chain-of-thought utterances, and do not necessarily support human comprehension. Key open questions are (1) how to produce high-fidelity reasoning traces that truly reflect the conceptual circuit enacted by the model, rather than merely a spurious post-hoc explanation, and (2) how to best represent these traces to users. This could potentially be done in much richer formats than text or static images, for example allowing the user to interactively "jump into" intermediate steps, provide feedback or alter processing strategies.

An Opportunity for Revisiting IR Literature Finally, the emergence of retrieval augmentation offers an opportunity to revisit classical IR literature on interaction paradigms for history keeping that had been proposed (and often rejected) in the context of list-based SERPs. Examples include breadcrumbs of intermediate search and reformulation steps taken by the user [150], and information scent and foraging theories [131].

4.2.6 Sources

There is a tension between retrieval and generation in RAG systems with relation to their sources. Retrieval ensures grounding in external sources, while generation produces fluent but sometimes unverified output. Abstractions such as embeddings or summaries simplify information but remove detail, and retrieval becomes essential for reconstructing these details. A key challenge is moving from abstraction back to specific information.

This same tension highlights how the information space is shaped by abstracting knowledge into models that organize and connect sources. It leads to important questions about the relationship between individual sources (i.e., databases, tools) and the larger information space they form together. It also raises the possibility that different kinds of information might give rise to distinct information spaces, which could be compared or combined.

The new paradigm of having a RAG system "read" the sources for us leads to numerous opportunities in enabling a wider range of tasks and users, but also poses risks that the community should take into account and discuss.

Wikipedia is where the search stops and RAG starts, is it enough? A study on conversational systems shows that many responses are drawn directly from Wikipedia articles [167]. This heavy dependence turns Wikipedia into the default external knowledge base for user queries while narrowing the scope of information and amplifying the gaps and biases in a single source. Retrieval-augmented generation systems follow the same principle of retrieving supporting evidence before producing responses. The reliance of conversational models on Wikipedia raises an important question for RAG research. Can Wikipedia be treated as a

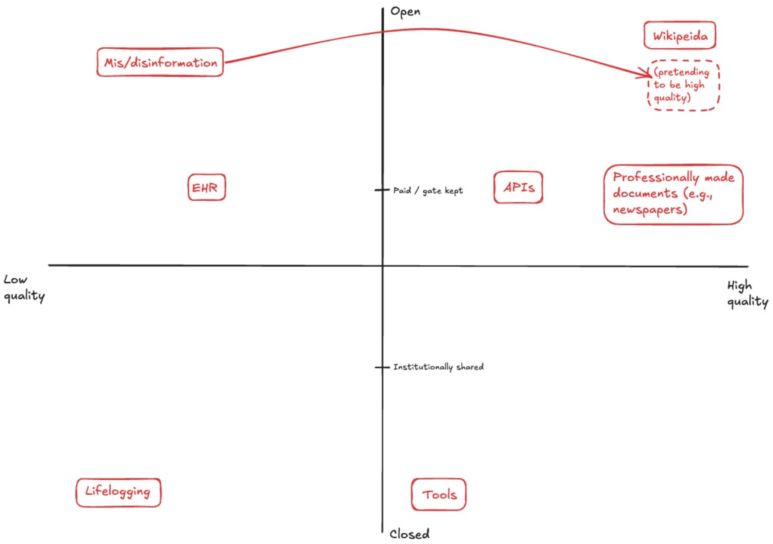


Figure 6 The source space, including familiar sources that we can envision fall in this categorization, as well as cases where the quality of the source is low, but the goal of the creator is to show it is high quality (e.g., mis/disinformation).

sufficient knowledge source, or must effective RAG include broader and more diverse corpora to overcome these limitations?

A large portion of user prompts and information needs can be addressed by relevant Wikipedia articles, but we envision a broader perspective where various and diverse sources can be leveraged by the system. Figure 6 depicts our vision of how different sources of information can be classified and leveraged by RAG systems, as well as the opportunities and challenges that come with them.

As seen in Figure 6, sources can be classified along multiple axes. One axis ranges from low to high quality, reflecting trustworthiness and accuracy. Another axis ranges from open to closed, covering open access, paid or restricted sources, and institutionally shared collections. Further axes could include modality (text, structured data, multimedia) and stability (static versus dynamically updated).

Figure 7 shows the information density to the user. In other words, it depicts the amount of information that is available to the user at different stages of RAG. As we see in the plot, one vision (the blue dotted line) is that as the user moves through the different RAG stages, the amount of information decreases significantly, because the system knows more about the user's information need. Clearly, as soon as the user enters a query and searches for it, the information space is much more limited. Moreover, a RAG system can provide a short summary or a tailored response to the user; therefore, the user would not need to go through all the results in the list. However, in another vision, the system could provide additional information that is not in the ranked list, as it knows more and more about the user's task. Therefore, the dashed green lines show the two possibilities at this stage where an efficient RAG system could leverage its internal knowledge, as well as the ranked documents, to provide more information or more steps in solving the task at hand.

Benefits. One key benefit of RAG systems is **consistency**. Placing a generative model at the end of the pipeline, information can be presented uniformly across multiple outputs. This consistency improves accessibility for diverse user groups, including elderly users and

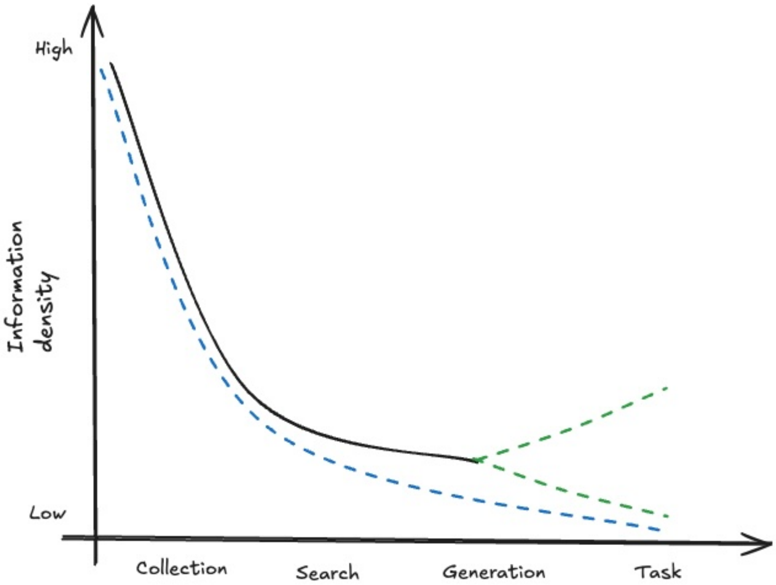


Figure 7 The information density diagram.

individuals with disabilities, by providing information through a uniform and consistent channel. Consistent results presentation also reduces confusion when users access the system repeatedly or compare outputs from different queries.

Another advantage is the **knowledge synthesis from multiple sources**. RAG systems consolidate and integrate information from diverse places, making information discovery more efficient and reducing the cognitive load on users. By synthesizing complex or dispersed information, these systems allow users to focus on higher-level analysis, decision-making, or learning. This benefit extends to a wide range of users, enabling them to access insights that might otherwise require significant time and expertise to assemble.

RAG systems can also enhance relevance and personalization. Combining retrieval and generation allows the system to prioritize information that aligns with the user's needs, preferences, or context. This tailored approach improves the efficiency and usefulness of information access, supporting tasks ranging from research to everyday decision-making. In addition, RAG systems support critical reasoning and exploration. By exposing users to synthesized knowledge drawn from multiple sources, they enable cross-referencing, comparison of perspectives, and identification of patterns that might not be obvious from a single source. This capability can promote more informed and balanced understanding, particularly in complex or rapidly evolving domains.

RAG systems can also facilitate scalability and adaptability. For instance, they can process vast information, update outputs as new sources become available, and generate summaries or explanations in multiple formats or modalities. This flexibility makes RAG systems valuable across domains, user groups, and applications, from education and research to healthcare and public information services.

Risks. RAG systems introduce a number of risks that affect both the reliability of information (i.e., sources) and the way users interact with it. One concern is that sources' quality, reliability, and authority may be ignored or flattened, which undermines the trustworthiness of outputs and can amplify existing biases. Another problem is the explainability paradox. That means the presence of citations creates the impression of transparency and explanation, even when the connection between the cited source and the generated text is weak or misleading.

Erosion of Credibility: Credibility and provenance are also at risk. The tone and style of original sources are often lost, and diverse voices are merged into a unified generated output. This process erases the distinct identity of sources, diminishes credibility, and makes verification more difficult. Alongside this, the potential for misinformation and disinformation is high. Generated text may contain inaccuracies, and since users often do not fact-check, quality differences are hidden behind the smooth surface of generation.

Another risk is epistemic homogenization. By blending diverse perspectives into a single synthesized response, RAG systems can erase differences across sources and reduce the visibility of plural or contested views. Simultaneously, fluent language encourages users to overtrust the system, fostering cognitive offloading, weakening information literacy, and discouraging verification against original materials.

Temporal and contextual risks complicate the use of retrieved sources. Content may be outdated, incomplete, or detached from its original setting, and generation can obscure these issues by presenting the information as current and authoritative. This masking effect increases the likelihood of misinformation and disinformation, since users often do not verify or fact-check what they receive. The absence of transparent provenance makes it even harder to distinguish reliable material from misleading or false claims. Feedback loops introduce additional systemic risk. Minor inaccuracies and distortions can accumulate and amplify when generated outputs are recycled into retrieval pipelines or incorporated into training data. Over time, this process undermines credibility and accuracy and may contribute to model collapse in extreme cases. In addition, this introduces a wide variety of risks of misuse of the generated content (see more in Section 4.4).

Opportunities. // Research questions. **Expanding RAG Beyond Text.** Advantages in LLMs and multimodal language models enable RAG systems to move beyond textual information. RAG systems can synthesize richer knowledge and provide more comprehensive responses by leveraging images, audio, video, and other modalities. The following research questions aim to explore how multimodal sources can be effectively integrated, what challenges arise, and how these approaches reshape the landscape of information retrieval and generation.

- How can RAG systems be designed to incorporate multimodal information beyond text?
- How should RAG systems present results that integrate multiple modalities to ensure clarity, usability, and effective communication of synthesized knowledge?
- What methods are needed to enable LLMs and multimodal language models to retrieve, align, and synthesize knowledge across multiple modalities?
- How does integrating multimodal sources change the structure or representation of information spaces within RAG systems?
- What opportunities and risks emerge when RAG systems move beyond text, particularly regarding accuracy, bias, interpretability, and user trust?

Personal Information Management. RAG systems have the potential to access a broad range of private and personal information, as illustrated in Figure 6. This access opens significant opportunities for personal information management, such as developing intelligent personal assistants or lifelogging systems, a long-standing vision in the information retrieval community. The reasoning and synthesis capabilities of LLMs applied to heterogeneous personal data create new avenues for research, particularly around how these systems can manage, interpret, and utilize sensitive information effectively.

- How can RAG systems effectively manage and synthesize diverse personal and private information to support intelligent personal assistants?
- What methods enable LLMs to reason over heterogeneous personal data while preserving privacy and security?
- What are the ethical, privacy, and trust considerations when RAG systems access and process sensitive personal data?
- How can results from personal sources be presented in an interpretable, actionable, and user-friendly way for lifelogging or personal assistant applications?

Tools, Databases, and APIs. RAG systems are increasingly capable of interacting with external tools (i.e., applications such as email, calendar, private repositories), databases, and APIs. Existing generative AI systems already try to leverage different sources to solve tasks by combining multiple types of information. In the context of RAG, different sources can provide complementary services, such as recommendation tools offering personalized suggestions or search engines supplying factual information.

- How should RAG systems present combined results to users?
- How can LLMs combine outputs from different sources accurately?
- How can RAG systems choose the best source or tool for a specific query?
- What problems arise when coordinating different types of services, like recommendations and factual data?

Quality-driven Real-Time Bidding (RTB). In a world filled with LLMs, each of which having its own expertise, we envision a competitive environment of various RAG systems aiming to contribute and be part of the user experience. This could lead to a quality-driven RTB, where LLMs can bid on the next token that goes as the system output. In other words, a quality-driven bidding system (e.g., based on the LLM's uncertainty, predicted performance, etc.) can determine which LLM gets to generate output for a given number of tokens, enabling a collaborative environment of LLMs. Therefore, we envision the following research questions:

- How would an ecosystem of a quality-driven RTB be designed to allow for fair and effective content generation?
- What could be the quality measures we could consider to rank the RAG systems in RTB?
- What would be the granularity of the RTB (e.g., token-level, passage-level, etc.), and how would that affect the overall quality and user experience?

Collaborating and Sharing with RAG Systems. As RAG systems become more widespread, scenarios become more prevalent in which users share or exchange their RAG systems or outputs. Such exchanges enable collaborative tasks, joint problem solving, or discussions grounded in shared information. Exploring how RAG systems can support collaboration and knowledge sharing raises several important research questions.

- How can RAG systems support effective collaboration among multiple users?
- How can shared outputs be presented in a way that is clear, actionable, and easy to understand? (See 4.2.5)
- How can user interfaces and interaction design be optimized to make collaborative RAG systems intuitive and efficient for multiple users?

Live Fact-Checking of Sources. While having the risks of generation disguising mis-/disinformation and giving the wrong impression of credibility of the user, LLMs can also help users in live fact-checking of the sources used when generating a response. This can be done in various dimensions, such as checking for the attribution of the generated content, checking the credibility of the cited sources, and checking the alignment of the final response with the utilized sources. We argue that the fact-checker LLMs have to be independent and different from the LLM used in the RAG system to provide a reasonably reliable assessment of the credibility of the sources. This could be integrated in the UI, in a similar way proposed in Section 4.5.

GAR. One approach to improving the user experience is to enhance ranked list presentation by combining retrieval with generation. In a GAR-based interface, search results could be enriched in multiple ways, such as fact-checking documents, generating more informative summaries, and incorporating images or other media. The interface could support interactive exploration and filtering, adapt to user intent, and explain why content is ranked or generated. Systems could learn from user feedback over time (i.e., user model), improve accuracy, and highlight potential biases to build trust. These features would enable more personalized, reliable, and actionable search experiences for the future.

- How can ranked lists use generated content to make search results clearer and more engaging?
- How can interactive and adaptive features help users explore and filter results in real time?
- What methods can GAR introduce to detect and reduce bias in retrieved documents and generated summaries?

4.3 Retrieval-Augmented Generation: The System's Perspective

Qinqyao Ai (Tsinghua University, CN)
Avishek Anand (Delft University, NL)
Michael Granitzer (University of Passau, DE)
Faegheh Hasibi (Radboud Universtiy, NL)
Djoerd Hiemstra (Radboud University, NL)
Sean MacAvaney (University of Glasgow, UK)
Arjen P. de Vries (Radboud Universtiy, NL)
Guido Zuccon (Google Research Australia and The University of Queensland, AU)

4.3.1 From Naive to Ideal: RAG System Architectures

The original RAG system architecture provided a simple, static pipeline from retrieval to generation—referred to here as *naive RAG*. In this setup, retrieval results are appended to the input of a large language model, which generates a response. More recent advancements have given rise to *active* or *dynamic* RAG systems, in which the generation module may initiate additional retrieval steps, for instance to reduce uncertainty during generation. In what follows, we articulate the vision for an *ideal* RAG system, then contrast it with the naive baseline in Table 1.

An Ideal RAG system

We envision an ideal RAG system that is a **compound AI system** that learns, reasons, and communicates, **continually evolving** and **adapting** to users, data, and the world itself, while **making principled trade-offs** between effectiveness, efficiency, and cost.

Limitations of current RAG Systems

Current RAG systems are primarily designed as plug-and-play augmentations for LLMs, retrofitted onto existing search infrastructures originally built for human users. In these architectures, retrieval is treated as a basic, static component, while the focus in innovation lies on the generation components as the dominant one to optimize. Retrieval typically relies on keyword or dense methods over static, pre-chunked documents, with a fixed scope that is unaware of query semantics, task complexity, or downstream usage, lacking any adaptive feedback loop.

Reasoning capabilities are limited, often reduced to data-driven shallow reasoning like chain-of-thought prompting or its modern counterparts [151, 171, 178] resulting in overthinking and inefficient unnecessary sampling path.

These systems are also poorly equipped to handle evolving knowledge, and feedback [136]. Updates typically require retraining or full re-indexing, with little support for continual or incremental learning. Furthermore, resource usage is uniform and over-provisioned, allocating the same compute regardless of query ambiguity or difficulty [159]. Uncertainty, whether in retrieval coverage or generation confidence, seldom surfaced. Users are expected to trust outputs without access to provenance, citations, or error bounds [147, 152]. Finally, most current RAG deployments rely on centralized architectures, retrieving from global indices governed by the provider. Users have minimal, if any, control over what data is indexed or retained. System evaluation is narrowly focused on effectiveness metrics like accuracy or relevance, with little consideration for energy, latency, or other operational costs.

Characteristics of Ideal RAG Systems

- **Adaptive Efficiency-Aware Execution**
  - Dynamically adjusts compute usage based on task complexity.
  - Optimizes compute vs. effectiveness tradeoffs.
  - Predicts and operates at the minimum resource footprint needed.
- **Flexible Memory and Communication Channels**
  - Uses multiple memory types: short-term, long-term, editable.
  - Supports communication via text, vectors, or symbolic formats.
- **Uncertainty-Aware Design**
  - Tracks uncertainty from user intent, knowledge, and model confidence.
  - Surfaces and responds to uncertainty (e.g., via clarifying questions).
- **Federated and Decentralized Architecture**
  - Supports domain-specific retrievers and data ownership.
  - Adheres to privacy, policy, and regulatory constraints.
- **Composable, Modular Components**
  - Enables independent optimization and replacement of components.
  - Easier to update and maintain with minimal retraining.
- **Cost-Aware and Sustainable**
  - Operates within latency, FLOPs, or carbon budgets.
  - Emphasizes caching and hardware-aware execution.
- **Feedback Aware**
  - Supports feedbacks from different parts of the pipeline.
  - Allows different types of feedback from scalar, descriptive, and is able to adapt to feedback.
- **Multi-Level Reasoning and Planning**
  - Supports contextual reasoning and task-aware planning.
  - Allows modular integration of symbolic, analogical, or causal logic.

Comparing Naive and Ideal RAG Architectures

4.3.2 LLMs as Search Engine Users

From the onset, search engines have been designed for human users. Consequently, retrieval models and evaluation measures have been designed around human-centered notions of relevance and human behavior, and search systems have been optimized accordingly. This means for example that queries are typically issued in natural language one-at-time and results are typically presented as ranked lists intended to be examined by the user in sequence. These same search systems are now central to RAG, where a retriever supplements the input of a large language model with external documents. Thus, retrieval in RAG relies on methods optimized for humans rather than for LLMs. However, recent empirical studies have suggested that search results as returned by search engines might impair LLMs. For example, a model may prefer documents not to be ordered by relevance. Instead, it might even perform better when non-relevant documents appear first in the context. These findings raise a deeper question: How would a search engine for LLM be different from a search engine for humans?

Table 1 Comparison between Naive and Ideal RAG Systems

Aspect	Naive RAG Systems	Ideal RAG Systems
LLM Interaction	Search designed for human consumption	Interfaces optimized for both humans and LLMs
Efficiency	Static compute for all queries	Adaptive, task-sensitive compute allocation
Reasoning	Shallow, post hoc reranking	Rich, integrated multi-level reasoning
Uncertainty	Hidden or implicit	Tracked, surfaced, and actionable
Architecture	Centralized monolith	Composable, federated modules
Memory	Flat, fixed-length context window	Hierarchical and type-specific memory structures
User Control	Minimal to none	Full control over indexing and retention
Evaluation Focus	Effectiveness only (e.g., accuracy)	Multi-objective (accuracy, efficiency, cost)
Maintainability	Costly updates and retraining	Incremental updates, easy maintenance
Feedback	No Feedback	Active & granular feedback

We envision **an ideal RAG system** as one in which the information needs of the generator can be flawlessly expressed to the retriever, and the retriever flawlessly expresses exactly the required information to the generator. In other words, there is flawless coordination between the retriever and the generator. This might mean having to renegotiate how information is represented and exchanged between the search engine and the LLM, how search strategies are planned and executed, and even re-thinking the overall architecture of RAG and the separation between retrieval and generation capabilities [98, 163].

Existing approaches have not yet achieved this ideal. Systems that "memorize" information in the generator's *parametric memory* suffer from forgetting [149], difficulty in attribution [28], and lack an expandable memory store [177]. Meanwhile, systems that invoke *Search Engines as a Tool* primarily use natural language as an interface between the LLM and the search engine, which is inherently ambiguous: natural language queries of a finite length cannot fully provide all context surrounding the information need, and results expressed in natural language lose their context.

To move towards this ideal system, we propose the following research directions:

- **Better understand LLM "behavior":** We argue that as understanding (human) user behavior has allowed the IR community to refine and optimize search systems to better support users, so understanding LLM behavior will allow us to better design search systems for LLMs. For this, empirical observations and methods from Machine Psychology [67], which aims to use theory and practice from human psychology to better understand LLM behavior, might help us determine what are the factors that influence RAG, what these mean for search engine design, and how LLM behavior in this context can be steered or controlled.

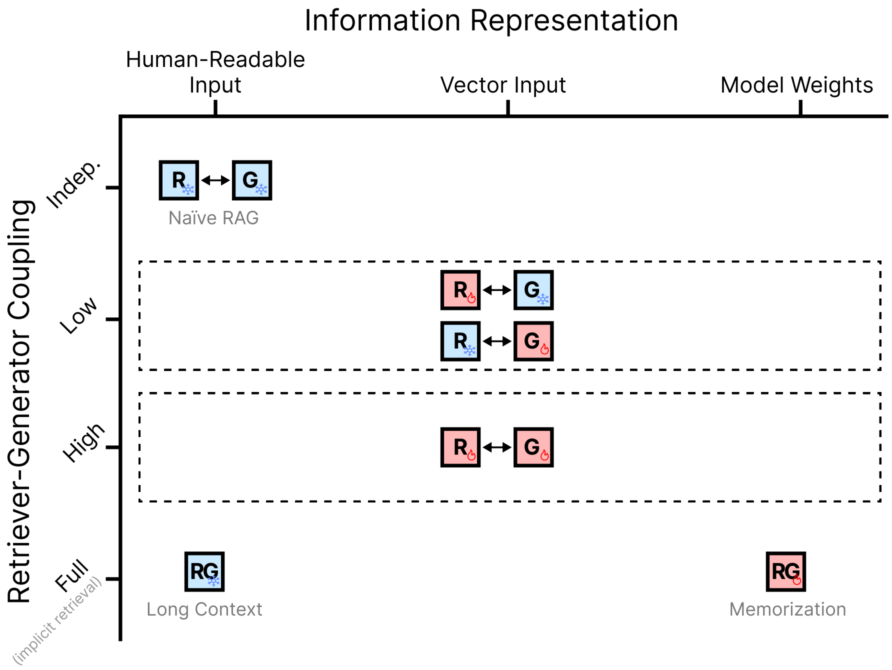


Figure 8 Comparison of current RAG approaches with respect to the representation of information and the coupling between the [R]etriever and [G]enerator components. A red box indicates the component has been trained for the task/data; a blue component indicates it is frozen (i.e. not specifically trained for the task/data). Naïve RAG interfaces between distinct retriever and generator components through human-readable text. Combined Retriever-Generators ([RG]) have been explored in long context settings [97] and through direct memorization into model parameters [164].

- **Enhancing Text Interface:** Differences between LLM and human user behaviors present opportunities to better align the text interface between the retriever and generator. These could be changes to the input, output, or interaction interfaces. On the input interface side, we could design a new query interface that aligns better with the way generators express their information needs (rather than the keyword or natural-language queries that current retrieval systems use). Perhaps this would mean supporting semi-structured queries to allow the generator to specify very specific information needs (which are challenging to express precisely in natural language and challenging for human users to write). In terms of the output interface, we could design new ways to present the retrieval results instead of a ranked list of documents. Perhaps this could involve technical details of the retrieval process, which are challenging for human users to interpret but could help the generator understand whether any follow-up requests are necessary. In terms of generator-retriever interaction, suggestions or clarifications to the query could be requested before fully executing the query. This could be inconvenient for a human user to use, but could help control retrieval costs and ensure that model context tokens are used effectively.
- **End-to-End Optimization:** Retrievers are typically optimized only for relevance to queries; within RAG these retrievers may surface information that is redundant, topically relevant but unhelpful for grounding answers, or even misleading (e.g. if it gets the generator to anchor to noisy evidence). Jointly training the retriever and the generator under a shared, task-specific objective promises to align retrieval with what actually helps

the generator produce correct, faithful and useful answer. This end-to-end optimization can involve gradient-based methods (if the retriever is differentiable, backpropagate through generator), policy gradient/reinforcement learning methods (treat the retriever as an action policy, optimize the reward based on the generator outputs as assessed through correctness, faithfulness or user feedback), and labels (by generating supervision signals that tie retrieved content to the final answer quality, most likely synthetically). However, jointly training the retriever and the generator presents a number of challenges. Generators are optimized with cross-entropy on tokens, not on truthfulness or factual grounding, therefore raising a training signal mismatch. This makes it hard to propagate useful gradients back to the retriever, especially when the reward is a "sparse" judgement, e.g. the final answer was correct. On the other hand, context length and inference cost constraints limit how much can be retrieved and passed as input to the generator. Optimising not just what is retrieved, but also how much and in what form (long documents vs. chunks vs. summaries/snippets) is still an open question; compression, summarisation and structured retrieval (tables, knowledge graphs) further complicate end-to-end training. Another open direction is how to encode feedback for end-to-end optimization, and in particular negative signals: if an answer is wrong, was it becausse the retriever pulled irrelevant documents, or because the generator failed to "reason"? Disentangling responsibility between retriever and generator is tricky for optimization. Similarly, current automatic metrics (e.g., token F1, BLEU, ROUGE, nDCG for retrieval) do not capture the nuanced success criteria of RAG (faithfulness, factual correctness, attribution, reasoning quality). Without good metrics, however, it is hard to optimize reliably end-to-end. Scalability and efficiency aspects of joint training also should not be underestimated: this training is computationally heavy, especially with large corpora and long-context models. Promising directions for exploration include differentiable retrieval mechanisms to enable smoother gradient flow; RLHF-like training for RAG, where human feedback (or synthetic judges) rewards correct and grounded answers; adaptive retrieval models that decide when/how much to retrieve, rather than always retrieved k documents; and the creation of unified architectures that tightly couple retrieval and generation, e.g., retrieval-augmented transformers where retrieval is embedded in the attention mechanism.

- **Representation of information:** The retrieval component is responsible to return information to the LLM to ground generation. Most common systems currently return this information as a rank list of search results, where each result is represented by the text contained in the document (or chunk or snippet). However, alternatives are possible, and representations could be defined across a spectrum that goes from human-readable text (the content of the document, snippets from the document, summaries of documents, etc), to embedding-based representations (e.g. as done in prompt compression approaches [78]) and model weights [160]. The use of model weights as a communication channel for RAG offers the opportunity for more principled and close-coupling approaches, but raise challenges such as how retrieval knowledge can be increased without forgetting, and how attribution can take place.

4.3.3 Uncertainty in RAG

Uncertainty is an important factor to consider when consuming information. Knowledge about the degree of uncertainty associated with that information is essential to make informed decisions and to identify potential risks in advance. Studies on uncertainty in IR have mostly analyzed the reliability of information sources and attempted to increase our understanding of the behavior of retrieval systems [64]. The introduction of generative AI systems in

RAG complicates matters, as some degree of uncertainty is also inherent to their behaviors and outputs [130, 152]. Therefore, understanding, modeling, evaluating, and expressing uncertainty is an important part of our discussion on RAG from the perspective of the system.

In the following paragraphs, we briefly describe how to view the concept of uncertainty in RAG and what we believe to be fruitful research directions in this field. Specifically, we first introduce what we believe an ideal system would do with uncertainty and how current systems perform so far, then we discuss what are the potential steps to guide us to achieve our goals.

Where we want to go

Uncertainty may or may not be part of the final output users expect from a RAG system, but it is undoubtedly a key piece of information that can help improve system performance and support downstream applications; e.g., abstaining from providing low-confidence outputs [57, 152]. Ideally, a RAG system with good support for uncertainty quantification should have the following abilities or characteristics.

First, an ideal RAG system should be able to understand and track any sources of information or reasoning uncertainty in its working pipeline. That means that when anything unexpected happens, we could trace back or at least evaluate which part of the system could be wrong. In practice, uncertainty could come from many places. Examples include but are not limited to (1) users, who provide vague information request, do not have a clear image on what they are looking for, or even may not possess the skill or expertise required to understand the system's outputs; (2) systems, which have multiple modules that work together to create the final output, and each module has their own reasons to introduce uncertain results, e.g., retrieval modules could be affected by documents from sources which have uncertain reliability, LLM could be affected by the bias, knowledge conflicts, or noise in its training data, etc.; (3) the world itself, where the information environment is dynamic and the provenance of each piece of knowledge and information may not be fully traceable and verifiable.

Next, with a good understanding of the uncertainty sources, the system should also have the ability to accurately quantify the uncertainty contributed by each module, e.g., uncertainty in the retrieval results, response generation process, and query reformulation. This not only provides important signals for debugging, but can also help us further improve the quality of each part of the system accordingly. For example, if we notice that the final output is unreliable mainly because of unreliable retrieved documents, then we could apply a more sophisticated filter and remove low-quality documents first. Finally, when presenting the output to real users, the system should have an effective way, perhaps a well-designed UI or data visualization tools, to express its uncertainty to users so that they can be fully informed about any risks behind the system and make wise decisions accordingly.

Where we are now

Many studies in the literature have analyzed the uncertainty of machine learning systems [35, 44]. Within the field of IR and NLP, there is also progress in identifying the uncertainty of documents [84], retrieval models [137], and LLM [45, 147]. For example, IR researchers have tried different methods to analyze the reliability or authority of documents on the Web [127] and social media [191]. There is also work on how to calibrate the ranking scores produced by learning-to-rank systems to correctly reflect the probability of relevance of each retrieved document [37].

For LLMs, researchers have explored different methods to analyze the internal states of LLMs and their output confidence (mostly from the perspective of token-wise or sentence-wise perplexity) [17, 161, 187]. These methods, however, are suboptimal for the RAG setup [153], as they only consider LLMs as the source of uncertainty. For RAG systems that combine all these modules above, things are still at an early stage. There have been works that use the advantages of uncertainty quantification techniques to improve RAG systems [79, 159], but research that directly studies uncertainty in RAG is still limited in the literature. Few studies have explored uncertainty quantification in the standard retrieve-then-generate RAG framework [130, 153], or in more advanced retrieval-augmented reasoning settings [154], where LLMs employ search engines as tools during the reasoning process [80, 186]. There are many challenges that prevent us from identifying and quantifying uncertainty in RAG systems effectively and efficiently, some of which are discussed in the next section.

As for the expression of uncertainty to users, there are some efforts that try to present an uncertainty score together with the output of the generative system in text and visual presentation modes [95]. However, the uncertainty scores provided are usually not grounded, and the way how systems express uncertainty to the user is by no means comprehensive and perfect.

What to do next

To achieve our long-term goal towards a reliable RAG system with uncertainty support, we brainstormed and identified four important challenges and directions to work. Specifically, they are:

- **Identification and Control of Uncertainty Sources**. Finding the sources of uncertainty is the foundation of all downstream processes. As discussed above, uncertainty in RAG systems has diverse sources including users, systems, and the dynamic environments of the world, e.g., ambiguous user queries, incomplete retrieval coverage, inconsistencies in the training data, or knowledge updates to science or events that are time sensitive. Identifying uncertainty sources requires not only tracing the provenance of information, either retrieved or generated by LLMs, but also translating this provenance into measurable signals of reliability. This is particularly difficult when there is not a clear definition of the unit of information. For example, documents could be chunked into parts with various lengths. Also, the usefulness and reliability of a piece of information cannot be fully understood when the context is fragmented. We need to figure out a more formal taxonomy of uncertainty and its granularity so that we can better understand and analyze the sources of uncertainty.
- **Uncertainty Quantification**. Together with source identification, quantification of uncertainty is another challenging problem. Even if we know the provenance of uncertainty, translating this provenance into measurable signals is still difficult because there is no “ground truth” for it. For example, existing studies mostly measure the quality of uncertainty quantification based on their correlation with the actual correctness of the system’s output [16]. However, this cannot serve as rewards to train an uncertainty quantification model that covers different sources of uncertainty because that would be essentially training a model to fit the task itself and may not generalize to all cases. Unsupervised methods are the state-of-the-art for uncertainty quantification, but there is not a clear roadmap for how to keep pushing the limit of these methods in the long term. More theoretical studies are needed on the optimization and evaluation of uncertainty quantification [152].

- **Presentation of Uncertainty**. Presenting uncertainty to users introduces another layer of complexity. Systems must communicate uncertainty in a way that is understandable, actionable, and aligned with user needs. This could mean asking clarifying questions, showing supporting evidence, or designing simpler or more complicated user interfaces. Also, generation and retrieval behave differently with respect to how to show uncertainty. Retrieval makes it explicit through showing the documents and providing citations, while generation tends to minimize or mask it. This raises open questions about when citation is necessary (e.g., factual claims like "Paris is the capital of France" may not require attribution) and how to reconcile conflicting or incomplete information.
- **Utilization and Reduction of Uncertainty**. Assuming that we could get reasonable uncertainty quantification for RAG systems, how to use uncertainty to improve the system is also an open question. For example, uncertainty could be an important resource for model optimization, where it could help in filtering out biases, inconsistencies, and other types of noise in the corpus that may propagate into outputs. It could also serve as a signal to guide agentic systems to design their workflow automatically for tasks that involve exploration of multiple sources of information. The potential of uncertainty in RAG, and other agentic systems that have retrieval and generation components, is not fully explored.

4.3.4 Efficiency vs. effectiveness trade-offs

The generation part of retrieval-augmented generation (The G of RAG) requires, in practice, the use of pre-trained large language models (LLMs). LLMs are notorious for their need of special hardware (i.e., GPUs) for both training/fine-tuning and for running them in production. The costs of running LLMs are orders of magnitude higher than the costs of running a standard retrieval component (The R of RAG). The work group discussed various different costs, such as: Development costs, convenience/cognitive costs, latency costs, query throughput costs, training vs. inference costs, and energy costs. This subsection focuses on the energy costs of training/fine-tuning of research experiments and the energy costs of testing/querying as a focus of efficiency. Our discussions were guided by the two overarching questions: 1) What are best practices of measuring energy-usage of systems? and 2) What efficiency vs. effectiveness trade-off is reasonable? Concretely, we asked ourselves: What are realistic lower bounds on efficiency? When is a small increase in effectiveness not worth it, given a large decrease in efficiency? When is it – the effectiveness – enough, if ever, given both engineering limits and societal limits?

Best practices

Strubell et al. [158] describe a best practice for estimating energy-usage of LLMs, at both training and inference time. A best practice adapted for retrieval experiments by Scells et al. [140] shows that roughly at query time, a standard (BM25) baseline is about 1 order of magnitude more efficient than a standard (LambdaMart) learning-to-rank baseline, and a staggering 5 orders of magnitude more efficient than an LLM (monoBERT) rerankers.[4] Precise measurement of efficiency is hard, and may even need researchers to standardize on hardware [133]. Because the energy demands of systems differ by many orders of magnitudes, rough efficiency estimates suffice for informed trade-off decisions. An important challenge for

[4] The BM25 baseline likely becomes another 2 orders of magnitude more efficient if it retrieves a top 10 instead of a top 1000.

the research community is to start reporting the estimated energy-usage of their experiments and put the measured effectiveness in perspective given their costs. It should be easy to incorporate best practices for estimating the energy costs into research, for instance using standard libraries like CodeCarbon [39].

Reproducibility challenges

Reproducibility of research has mostly focused on measuring the effectiveness. From as early as the 1960's [34], researchers in information retrieval and library science have been carefully constructing open datasets and test collections that help researchers to reproduce results and advance the field. Best practices for research into the information retrieval component of RAG systems are followed in joint evaluation conferences like TREC [172], CLEF [58], NTCIR [82], and FIRE [61]. Similarly, there are benchmarks and leaderboards for LLMs that almost exclusively focus on their effectiveness, such as the Open LLM Leaderboard [122] and LiveBench [179]. Some LLMs, like Bloom, report the energy usage [108] as well as the data used to train the base model and the data used for instruction fine-tuning, following guidelines for describing datasets [62]. Such LLMs are better suited for reproducible open science than other popular "open" models like Llama 3 or 4, and a much better fit than using LLMs via APIs of, for instance, OpenAI or Google. The European Open Source AI Index provides a guide for choosing open LLMs along multiple criteria [102].[5]

4.3.5 Federated Search - Why now?

Federated search [24] addresses a key challenge in information retrieval: accessing distributed data under different ownership, often with controlled access and domain-specific ranking systems. In spite of past research on federated IR for specialized domains like digital libraries [60] and the successful adoption of federated architectures with protocols like OAI-PMH [93], centralized search systems prevail; with the market dominance of Google and Bing for Web search as a primary example. Given the rise of RAG, will the dominating architecture for future search systems remain centralized or is RAG leading us to an alternative, federated search infrastructure, that will be managed and orchestrated by a large number of parties?

Let us consider the key factors triggered by the recent development of LLMs that may lead retrieval-augmented generation systems away from a centralized system architecture:

- **Ownership and Privacy:** Unlike centralized systems, where a single provider controls the indexing and access to data, federated architectures distribute authority across multiple content owners, each retaining control over their own collections and access policies. For RAG specifically, a federated system architecture would ease the deployment of unified information access solutions that access Web data, enterprise data, and personal sources like email and calendar. This decentralization offers greater respect for intellectual property and institutional boundaries, but also introduces challenges: How to ensure consistent privacy guarantees across heterogeneous systems, and how to enforce access restrictions when retrieval is coordinated by third parties.
- **Data Quality and Domain-specific Curation:** Careful curation of the contents of the 'memory' that the retrieval component may access (traditionally referred to as the 'document collection'), translates directly into improved outcomes. The costs of curation

[5] https://osai-index.eu, last accessed September 25th, 2025.

can be shared with the community of users of that collection. We foresee this scenario as realistic in professional contexts (e.g., online communities that curate high quality health data[6]), but also for general interest purposes; consider for example topics like 'Karate in Germany' that are maintained in public Web directories like Curlie[7] (formerly ODP). Given the expected reduction in data quality and improved redundancy through generative AI, need for curation will increase, while curation costs can decrease due to improved tooling.

- **Efficiency and Ease of Deployment:** Retrieval systems have significantly benefited from the availability of open source embedding and re-ranking models as well as the rapid improvements of vector stores. Wide availability of open source solutions and LLMs has reduced the barrier in setting up RAG endpoints with good quality, democratizing access to this type of technology (leading to 'everyone and their mother developing RAG').
- **LLM-based Planning and Routing:** Shallow reasoning techniques in LLMs like chain-of-thought also enable complex search strategies, where multiple queries can be issued, perhaps reformulated based on analysis of query results, in a completely data-driven manner. Although this might be less optimal than formal reasoning methods, this *data-driven reasoning* is easy to integrate in the LLM inference process, and leverages the out of domain generalization capabilities of LLMs.
- **Tool Usage and Standardized Protocols:** Tool usage and the rise of agentic AI triggered the development of interoperability protocols like MCP [51]. While the standardization is largely driven in an ad-hoc manner through industry, adoption is quite high. Considering the planning and routing capabilities of LLMs, it becomes viable and easy to expose RAG systems as tools for LLMs via those protocols. LLMs may also drive the selection of resources in federated search [176].
- **Economic Incentives:** Today's web search engines serve a dual purpose: helping users discover information while also directing traffic to websites, content owners, and publishers. This creates a win-win-win scenario for all parties involved. However, the rise of RAG systems disrupts this mutually beneficial ecosystem. By aggregating and synthesizing information directly, RAG reduces the need for users to visit original sources. Worse still, much of the content used to train LLMs was ingested without fair compensation to publishers. As a result, publishers and content owners now have little economic incentive to allow centralized RAG-based search engines to index and utilize their material. In contrast, a federated RAG system empowers them to retain control over their content while creating new business opportunities. By offering RAG or MCP endpoints, publishers can monetize access to their data and enable LLMs (or other systems) to consume their content—on their own terms.

Aside this list of potential benefits, several factors speak against a federated RAG systems. We see the following potential points against federated RAG systems replacing centralized search:

- **Efficient LLM Inference Infrastructures:** Large-scale centralized infrastructures can optimize inference efficiency through optimized hardware or optimized infrastructure setup, or may use energy sources that are more environmentally friendly. The same optimization might be harder to achieve for federated RAG systems.

[6] `https://medical-data-models.org/` [46]), last accessed September 25th, 2025.

[7] `https://curlie.org/`, last accessed September 25th, 2025

- **Limits of Planning and Routing:** Although LLMs have shown powerful planning and routing capabilities, there are upper limits depending on the number of components that can be deployed effectively. While central systems and LLMs can learn routing during training, federated RAG systems require inference time routing and planning, which could reduce the model's capabilities.
- **Increased Security Risks:** Federated systems per definition operate across system boundaries which makes them (i) more vulnerable to security threats and (ii) harder to protect.
- **Ease of Use of Central Systems:** End-users prefer a single point of access and smooth user-experience. It has to be shown that Federated RAG Systems can achieve the same user-experience as centralized systems and provide an efficient, effective and reliable service.
- **Data Mixing Benefits:** It has been shown that LLMs benefit from mixing training data from various domains [188], and can create relations between different information spaces. It is yet to be seen, whether this property can be retained in a federated system.

Derived Research Questions:

From this analysis we derive the following open research questions for Federated RAG Systems:

- How to do resource selection in federated RAG systems?
- What is the limit (if any) of the number of federated systems participating in a Federated RAG, considering query planning and overall performance?
- Are current protocols sufficient to ensure reliability and quality comparable to centralized systems?
- Will federated systems be more efficient and as effective as centralized systems?
- Could clean and curated data help to make federated RAG *more* effective than centralized systems?
- Can data mixing benefits of large, centralized models be compensated in federated RAG systems?
- How to ensure data privacy in the exchange between components of the federation?

4.4 Societal and Ethical Motivations for Inverting RAG to GAR

Johannes Kiesel (GESIS – Leibniz Institute for the Social Sciences, DE)
Bhaskar Mitra (Independent Researcher, Tiohtià:ke/Montréal, CA)
Josiane Mothe (INSPE, Université de Toulouse, UT2J, UMR5505 CNRS IRIT, FR)
Heather O'Brien (iSchool, University of British Columbia, CA)
Birte Platow (TU Dresden, ScaDS.AI, DE)
Stefan Voigt (Open Search Foundation, DE)

Positionality Statement

This chapter reflects the discussions and views of a limited group of individuals who worked within a limited time frame. The group responsible for this section of the report is diverse in gender and cultural background with expertise in computer science, social sciences, and humanities with library and information science and theology, but undoubtedly does not represent the rich, diverse perspectives of the full spectrum of people who are impacted by RAG technologies. We also acknowledge the lack of representation from other relevant disciplines and expert communities, who should inform and have an active say in how the relevant technologies are developed.

4.4.1 Introduction

Ever since humans collated written information in books or collections of books (libraries), searching for relevant pieces of information, reading, perceiving and 'understanding' has been part of human knowledge gain. Filtering and retrieving was always part of search and selection processes to identify those pieces of information that were required for further processing in information-access tasks. With the increased capacity and wide range use of large language models (LLM) that statistically generate text (and other modes of information) prompted in a written dialog with the user, the classic information retrieval and perception process is changed significantly. In particular, since generative systems are perceived as 'intelligent' by many human users.

Editors, publishers, library curators or search engine providers in the 'classical' information domains such as libraries, newspapers and the Internet, already had great responsibility to transparently, ethically and reliably structure, process and provide access to information; however, the development and massive use of LLMs has altered and condensed this process of information processing and provision significantly. Information artifacts are generated massively, in an ad-hoc manner and 'on the fly', while the original sources are often hardly referenced. A number of ethical, psychological, societal, legal, political and environmental concerns can be associated with this wide use of generative information access technologies. Retrieval-augmented generation, of course, can help to mitigate parts of these concerns, e.g., by incorporating and referencing sources in the generative process and by thus increasing transparency and accountability. However, this is still insufficient. Thus, we suggest a paradigm shift in current information access thinking, from retrieval-augmented generation (RAG) to generation-augmented retrieval (GAR).

While RAG focuses primarily on generating better answers through retrieved context, GAR emphasizes the retrieval process itself as the essential processing step, adding the essential transparency and accountability to information access.[8] In this new 'paradigm', the most critical component is no longer generation (G), but retrieval (R): sources and retrieval mechanisms become the transparent foundation of information access and processing, rather than a supplement.

In GAR, the LLMs serve as statistical-generative (dialog based) interface for exploring different information spaces and providing the output in the form the user wants it. This approach emphasizes controlled, systematic exploration of knowledge sources, positioning GAR as a pathway towards more sustainable and transparent information access systems that empower users to engage deeply with source materials rather than consuming ad-hoc statistically generated, non-transparent, but still appealing answers.

In this chapter we collect some key issues, concerns, and opportunities that can be associated at various dimensions and scales with the ubiquitous use of LLMs and RAG systems. From potential impact on individuals to impact on society and even on (geo)politics; from cultural representations of information to the rights of the creators of works of art or information artifacts. We by no means claim to address all socio-ethical dimensions of statistical-generative information access and processing in a complete manner. We report on the main concerns that were raised and discussed during this Dagstuhl seminar and contribute to a constructive discourse on the ethically and socially sustainable use of systems for accessing and processing information, combining approaches based on "retrieval" and "generation" in a meaningful way. Specifically, this chapter is framed by a discussion of knowledge, ethics, and human rights, and its implications for RAG vs. GAR. Next, we offer perspectives on several sociotechnical issues for GAR: scholarly communication, user cognition, emotional and mental well-being, democracy and political discourse, and language and culture.

4.4.2 Knowledge acquisition and ethics

The rise of new information processing modes in LLMs, agentic LLMs and RAG is a disruptive factor in how we store, process and present information. This implicitly raises the question of what information is per se, and what functions it performs for the individual, groups of people and in the overall social system – or rather which functions it should perform in the future.

Interrelation of information and knowledge. In the everyday language of the information society, information is often equated with knowledge. Like information, knowledge then is perceived as a free-flowing impersonal resource. In this perspective knowledge is "stored" traditionally in libraries, later turned into digital libraries, databases and, subsequently nowadays, in LLMs and RAG systems. The idea of "storing" and "collecting" knowledge in analogy to other resources is not incorrect but overlooks one fundamental difference: knowledge has a closer connection to us than other resources like coal or water, for instance. While these resources would still exist if humanity was wiped out, the existence of knowledge is depending on someone (or some group) who knows. As understandable as it may be in our post-Enlightenment, rational knowledge society to identify information with knowledge, it is nevertheless wrong. Information, facts, or data are only building blocks of knowledge that depend on an architect. Without such a counterpart, information is ultimately nothing more than ink marks, electronic marks or tokens. Conversely, however, the following also

[8] Our definition of GAR deviates from earlier use [12, 111], in that we include in GAR all information access systems that focus on retrieval, not only those systems in which retrieval is the last step.

applies: without access to information and the knowledge based on it, people are not able to shape their lives independently, responsibly, and freely, which touches on fundamental questions of being human.

Interrelation of knowledge and human rights. It is not without reason that the right to education always ranks high in international agreements (see UN Convention on the Rights of the Child, UN Charter for Sustainable Development Goals, etc.). Education is the path to knowledge. Knowledge is the basis for the ability to judge, decide, and act, and is therefore a prerequisite for participation. In this respect, knowledge is also a determining factor for freedom and independence. However, knowledge should not be seen here as a static resource that an individual or group of people simply "acquires" and then uses as they see fit according to their own needs. Knowledge and education should rather be seen as a continuous process through which the individual establishes a positive, dynamic balance in their relationship with the world. The process perspective emphasizes that the acquisition of knowledge cannot be limited to the product "knowledge". Rather, it is a constant state of mind that must be renewed again and again. Knowing that information and knowledge are specifically interrelated (see "interrelation of information and knowledge" above), the question now arises as to how continuity and activity, as core elements of knowledge and education, can be secured for people in the future when information retrieval systems and LLMs enter this process as agents.

What is up for discussion (and modes of discussion). Although ethics is characterized by great openness (as opposed to morality or values), ethical analyses are always location-specific in their consideration of specific constellations and dilemmas. Therefore, it should be clarified in advance which analytical approaches will form the basis of a consideration. Ethical reflection can be descriptive (descriptive ethics) or evaluative (normative ethics). The intentions (deontological ethics), goals (teleological ethics), or consequences (consequentialist ethics) can be the subject of consideration. If the focus is on the consequences, it must still be clarified whether this is done from an individual ethical perspective or a social ethical perspective, i.e., whether the starting point is the individual or groups of people or society as a whole.

In terms of the evaluative parameters of the analysis, ethical analyses are flexible in terms of specific topics, but are generally based on universal values such as human dignity and other general concepts. In contrast, an analysis based on moral aspects is more focused and concrete. By morality, we mean the entirety of norms that apply in a specific community. Morality is based on the insight that there are certain codes of conduct and judgment that enable constructive coexistence. Norms and conventions, on the other hand, are habit-based concretizations of morality. Laws are ultimately the institutionalized "final version" and highest form of concretization.

For the present analysis, an ethical approach is chosen that largely disregards moral perspectives. Furthermore, the possible consequences of LLM and RAG in knowledge systems are discussed, both from an individual and a social-ethical perspective (cf. micro, meso, and macro levels). Against the backdrop of such an analysis design, the following questions arise as ethically relevant issues:

Key questions. To what extent must architects who transform information into knowledge be human – both from an individual as well as from a societal point of view? What happens to the individual/to groups/society when humans play less of a role in the transformation of information into knowledge? And what does this mean for knowledge systems themselves? And finally: What dynamics might RAG systems in their various forms bring to the upcoming development?

Societal scale. As seen above: without a mind to access and transform information according to its individual circumstances, whatever is stored in libraries, databases or LLMs will remain unstructured and, in the long run, static. Therefore, even a comparatively small share of an individual is indispensable in the transformation process from the perspective of the system. However, it is unclear to what extent and to what degree this effect must be reflected from the perspective of the system. Is a prompt sufficient? A query? What sources should an individual person access and process in order to fulfill their function as an information transformer and knowledge architect? To put it bluntly: Should the individual user primarily receive synthetic sources from the LLM, or should they (based on traditional knowledge generation methods) perform source work and assemble these as building blocks with the help of an LLM that supports them as an interface?

From an individual and ethical micro perspective, freedom, self-determination, independence and participation manifest themselves as learning and the ability to learn, which must be upheld as fundamental anthropological constants, especially in times of generative AI. In this sense, information is not only a prerequisite for the diverse realization of human dignity, it must also prove itself in context and at the same time be understood as a means by which humans themselves continuously fulfill their constitution as learning entities. This implies necessarily intensive and process-emphasizing involvement with information beyond simple prompting.

As seen above, knowledge belongs to some individual but at the same time it belongs to particular groups. What is more: the knowledge of a group may go beyond the knowledge of its individual members. Therefore, it is crucial to consider the above-mentioned aspects in a differentiated manner in the context of space and time (meso perspective) as they manifest themselves differently depending on the particular cultural context.

Since knowledge always serves the purpose of establishing a positive reciprocal relationship between individuals, the social environment, and society as such, society as a whole must also be considered from a macro perspective. It is obvious that the different approaches of these groups to information leave their own mark, which again corresponds to the idea of successful architecture.

The shaping of information corpora is therefore carried out for good reason and with justification by various "architects". However, it remains unclear which tools should be provided to these architects so that they can perform their design work (from information to knowledge) and at the same time fulfill their aspirations (knowledge as access to the world).

RAG vs. GAR from an ethical perspective. First, it should be noted that the various forms of generative AI open up unprecedented opportunities for individuals, groups, and societies to acquire knowledge. Against this backdrop, it is to be expected that knowledge will potentially be strengthened in all its positive effects (participation, freedom, etc.). However, this expectation must also be viewed alongside the possibility of limitations or even negative effects.

In this context, the decisive factor is likely to be the way in which people participate in the transformative work that turns information into knowledge. An excess of generative processes would reduce the diversity of sources and positions and, as a result, orientation and the ability to tolerate ambiguity and orientation as complex knowledge bases for accessing the world. As a result, freedom and independence would also be at least potentially endangered. In addition, it would reduce human involvement insofar as the product of information processing would become dominant over the encounter process. Human learning would not be taken seriously, and the transformational work of humans would be reduced to mere consumption in the long run.

In contrast, retrieval strengthens the aspects mentioned above. The generative elements in the transformation process could support communication, profiling, and depth of the transformation process without, however, contributing to the actual function of transforming information into knowledge by creating a relationship between the enormous information base and a knowledgeable person. Generation would then no longer be the constitutive factor in this process (RAG), but rather the connotative mediator (GAR).

4.4.3 Scholarly Communication and Publishing.

The Internet and social media (e.g., Twitter/X, YouTube) have transformed how scholars engage with each other and non-academics, highlighting how topics (e.g., Covid-19), communication style, and digital platform and audience factors contribute to or deter engagement with scholarship [26, 27, 68]. At the same time, memory institutions like libraries have moved from catalogs to discovery systems that facilitate access to a range of in-house (e.g., institutional repositories, open access digitized collections) and proprietary materials purchased from publishers and vendors [20]. Some may even provide altmetrics, bringing mentions about the item on various social media platforms into item-level descriptions. This has resulted in a 'one stop shop' for users affiliated with an institution to locate and retrieve books, journal articles, archival and special collections, academic databases, etc..

Generative artificial intelligence (Generative AI) tools (e.g., ChatGPT, Google Gemini, Perplexity AI, Le Chat) are disrupting scholarship, bringing new opportunities and challenges to effective research dissemination, uptake and use. To date, research has focused on the ethical and transparent uses of Generative AI in research production, e.g., authorship, peer review, policies to curb academic misconduct [38, 71, 96]. Generative AI outputs have the propensity to produce "fake but convincing articles" [38] (p. 235), false citations, and biased outputs [49, 96], and to separate content from its sources [142], all of which could threaten the legitimacy of research [71] and its potential to impact society through, for instance, policy development, decision making, etc.

Shaping Scholarly Communication Systems and Products. There are many actors in scholarly communication and publishing: scholars (who also act as peer reviewers, editors, and mentors to the next generation of scholars), publishers, funders, GLAM institutions (galleries, libraries, archives, and museums), scholarly and professional associations, and commercial enterprises that make products to assist scholars in tasks like writing, e.g., Claude, Writefull, or provide current awareness and dissemination services, e.g., ResearchGate, Google Scholar. Each of these actors create, promote, and use different information systems and formats or genres e.g., journal article, abstract.

Scholarly information systems may be reinvented in the coming years (and, arguable, this is already happening). GAR has the potential to break down silos between standalone systems. Currently, the searcher must navigate their own library discovery system, go to a search engine like Google, or a known web-based repository, e.g., arXiv, to initiate a search. GAR systems could provide seamless access to relevant content held by multiple scholarly information systems and in multiple formats (e.g., books, journal articles, data sets, multimedia). However, this requires us to consider questions around: copyright legislation, which differs by geographical jurisdiction; authors' intellectual property rights, including works created at different points in time under different (non-GenAI) conditions; protections for cultural heritage materials that may be held by institutions but rightfully belong to other people or groups, e.g., stolen Indigenous artifacts (cf., [100, 101]); and the role of GLAM institutions, which currently operate as intermediaries in the scholarly ecosystem. *There are opportunities for GAR researchers to work with GLAM institutions and publishers to shape* *information access and negotiate fair, transparent use of collections.*

Scholarly formats or genres are constantly evolving (e.g., the shift from print to digital journals), but the history of scholarly communication demonstrates how the materiality of research (i.e., form and function) shapes social, disciplinary, and institutional norms and practices [22, 41]. Currently, there are tensions in how scholars from different disciplines are approaching the use of GenAI tools in education and research settings, e.g., policies around tool use on course syllabi, disclosure statements on conference submissions, and whether this use threatens research legitimacy and academic integrity. Traditional university promotion and tenure systems are entrenched in academic attribution, e.g., bibliometrics or ways to 'count' (and therefore legitimize) authors' contributions. Therefore, it is important for GAR systems to support the accurate attribution of ideas and scholarly contributions, and to ensure that research remains a conversation – not just between humans and GenAI tools – but among researchers and public audiences. Scholars must be able to defend and verify their claims, correct misinformation or challenge harmful interpretations of their work. Finally, it is predicted that "research will be published in a way that can be easily read by machines rather than humans" [38] (p. 236). *This raises questions about what will constitute scholarly products and how norms around their construction will be negotiated within and across disciplines, and with GAR tools.*

4.4.4 Cognitive Processes in Information Seeking

Information behavior comprises information needs, seeking and use. Information retrieval systems are designed to assist users negotiate information needs, refine queries and search strategies, and assess the relevance of the results ("10 blue links") through various affordances, e.g., controlled vocabulary, search histories, links to related, relevant content, etc. [180]. Information systems are designed to support users to complete simple, factual and more complex tasks to accomplish a goal. Information behavior recognizes that affect, behavior, and cognition play a role in information interaction, and that these are intertwined. For instance, newspaper headings grab our attention (cognition) by triggering our emotions (affect), causing us to click and read (behavior). This section looks specifically at cognition.

Cognition in information behavior refers broadly to how people perceive and attend to, learn, process, organize and store information. Information systems require cognitive skills to use; for example, finding a document saved to your computer requires you to remember details like its title, key words, or location to facilitate retrieval. In other cases, information systems can compensate for limited cognitive capacity; for instance, a well designed product menu and facets can help searchers recognize and narrow down what they are looking for. There is no 'typical' user (though systems are often designed as though this is the case). Cognitive abilities are highly varied, with factors such as developmental stage, reading level, environment, neurodiversity, and others playing a role.

Generative AI systems are raising alarms for potential negative impacts on human cognition. Cognitive offloading refers to using tools like note pads, Google maps, or calculators to reduce our cognitive load [63]. Offloading to a Generative AI tool could be positive; if a user delegates some tasks to a tool, they may make space for higher order problem solving. However, studies suggest the opposite is true. Gerlich [63], in their survey of 600+ UK residents over the age of 17, observed a link between trust in Generative AI systems, cognitive offloading and reduced critical thinking. Another study by [90] recruited 50+ university students (aged 18-39) to be part of one of three groups: LLM, search engine, or brain-only. Participants' brain activity was monitored during three sessions where they wrote essays, either utilizing LLM or search tools or not; the essays were evaluated by two teachers using specific metrics. Based on the electroencephalography activities, the researchers concluded that the brain-only group showed more neural connectivity, indicating the utilization of more cognitive resources for the task, and more integrative activities at low frequencies, “possibly reflecting deeper encoding of context and an ongoing integration of non-verbal memory and emotional content into their writing” (p. 86).

These two studies represent emerging work, and more empirical work is needed to understand how Generative AI tools are changing our cognitive processes. Are these changes perpetuating cognitive decline or passive, uncritical information reception that could increase vulnerabilities for misinformation? Or do they present opportunities to build cognitive skills in key areas. The traditional search process consists of formulating the search, selecting sources, and interacting with sources [170]. If GAR systems ease the cognitive load of formulating the search and selecting sources, *what possibilities are there for prompting deeper, more critical engagement with sources if we move away from presenting outputs as “answers”? How can educators and information professionals help people to learn cognitive skills to prepare them to use GAR systems?*

4.4.5 Emotional and Mental Well-Being

As a relatively new technology, the long-term impact of LLM-powered conversational systems on individual’s emotional and mental health has not been studied adequately. New research is currently emerging that calls for much more measured deployment and adoption of this technology till its effect on mental health is better understood. The concerns here are not just about the generative AI itself, but also about how for-profit institutions may specifically train these models to maximize usage—*e.g.*, through application of sycophancy [145], anthropomorphic behavior [31], and emotional manipulation [40]. There are serious emerging concerns about these systems contributing to digital addiction [5, 112, 148, 185, 196] and even causing paranoia and delusions that has come to be termed as “AI psychosis” [48, 56, 89, 107]. It is particularly concerning that several cases of self-harm [14, 77, 88, 162, 184, 189] and accidental death [70] have been linked to chatbot usage. Reports are also emerging on cases of chatbot usage linked to negative impacts on personal and romantic relationships [47]. It is imperative that extensive studies are urgently performed to better understand the impact of these technologies on people’s emotional and mental well-being as well their interpersonal relationships, and that the technology is better regulated in light of those findings.

4.4.6 Democracy and Political Discourse

Online information access plays an important role in our collective sense-making of our place and relationships in this world, and mediates critical political discourse in society. We must thoughtfully consider how the applications of RAG for information access may shape the future of democracy. While these technologies may have much to offer in supporting future democratic processes, we must also critically examine the potential risks. We can already observe some of these concerns being reflected in public opinion in the context of the undue influence that chatbots are exerting to shape political narratives [85, 105, 123] and with respect to the risks of eroding public trust in democracy when chatbots are deployed irresponsibly by elected government representatives [110, 181]. Concerns about the impact of RAG (and more broadly generative AI) on democracy and political discourse includes misinformation and confirmation bias, AI persuasion, reduced transparency, and dependency on provides that we discuss next.

Misinformation and confirmation bias. Generative AI models are prone to producing factual inaccuracies in their outputs that may contribute to public misinformation [195]. This is further exacerbated by the fact that the application of generative AI models for generating concise summaries of information from retrieved artifacts shifts the responsibility of inspecting the information in the documents and assessing their relevance, trustworthiness, and surrounding context from the users to the AI models and disincentivizes users from developing critical cognitive skills necessary to distinguish between trustworthy and untrustworthy information [116].

AI sycophancy [145] may also encourage confirmation bias in users and lock them in echo chambers [146]. Democracy requires that all citizens, including those with opposing values, have access to the same shared facts. Otherwise, this may contribute towards further social decohesion[9].

Application of RAG for information access must be accompanied by appropriate safeguards to ensure that citizens have access to trustworthy information and a shared basis of facts, and do not negatively impact public information literacy.

Persuasion and Manipulation. Technical progress in approaches for aligning generative AI models towards specific values have been crucial in reducing harmful outputs. However, the same mechanisms may also be abused by system owners by combining them with massive amounts of user behavior data from surveillance capitalism [198] and generative AI's capability to produce persuasive language and visualizations [23, 25, 52, 116, 129] to censor and manipulate public opinion [85, 105, 123]. Such concentration of power to influence public discourse can pose serious threats to democratic processes. To address such risks, we must develop governance and auditing frameworks that ensure that the development and deployment of these technologies are performed under appropriate democratic oversight and are in alignment with our democratic goals.

We must also ensure that the application of RAG for information access does not further exacerbate user surveillance by encouraging users to share more personal information when engaging these systems in conversations.

Reduced Transparency. Democracy is based on transparent decision-making as a prerequisite for holding decision-makers accountable for their decisions. However, current LLM systems are not transparent and are not accountable for their results. At the individual level, citizens cannot trace where the information they use comes from, making them more susceptible to falling for misinformation. At the societal level, decision-makers cannot base their decisions on the results of opaque and unaccountable LLMs without violating the principle of transparency. Information access systems that focus on retrieval can mitigate this problem to some extent by providing users with results and their sources on which they can base their decisions or opinions on, rather than providing direct answers.

Dependency on AI Providers. Integrating a tool into workflows comes with a certain degree of dependence on that tool. In decision-making processes, such dependence can make people more susceptible to manipulation and cognitive biases (see above), as the incentive to rely on the tool is greater. Lock-in effects, such as the inability to transfer ones chat history from one LLM to another, exacerbate this risk, as users become dependent even on specific providers rather than on the technology in general. At the individual level, citizens may stick with (certain) LLMs even though they are aware of better alternatives. At the societal level, decision-makers may become dependent on LLM providers, giving them undue power

[9] We intentionally do not frame this as increasing polarization as that framing implicitly assumes that different sides of political discourse have equal merit which amounts to algorithmic bothsidism [115].

over the decision-making process. Information access systems that focus on retrieval can mitigate this problem to some extent, as they do not take over the process of forming an opinion or decision (but merely support it).

At the heart of these questions, lies the recognition of the fact that technological frames like RAG and GAR—as well as the decision to adopt one over the other—are all saliently political. To responsibly engage these technologies in our democratic processes and political discourse necessitates that we in turn also open up the sociotechnical imaginaries [115] that motivate their development as well as who gets to shape their realization also to social deliberation and democratic critique.

4.4.7 Cultural Representations

Worldwide vs Language Centric Information Access. On the search perspective, on traditional search engines, retrieval is mainly language-based: a query in French retrieves URLs / documents from resources in French. This is because of the principle of keyword search, where the query words have to match the document words, creating linguistic silos. This limitation is addressed by the architectures that combine search and LLM capabilities.

On the retrieval side, the frontiers between languages can be blurred thanks to the semantic representation. GAR offers the opportunity to retrieve information in different languages. Multilingual LLMs embed information in different languages and in Europe, the words "democracy", "démocratie" (FR), "democracia" (ES), "democrazia" (IT), "Demokratie" (DE), "democratie" (NL), "democracia" (OC) are supposed to cover the same concept of democracy for example. On the generation side, the user can get an output in a language different from the language of the retrieved sources/documents.

There is also the opportunity to have access to resources from different cultures through different languages [134]. Multilingual LLMs embed information from different perspectives [3]. As an example, Chinese medicine could become much more accessible for Westerners because the different pharmacopoeias share the common goal of providing solutions to human ailments for example. On the other hand, LLMs are not culturally neutral [86, 118]. The implicit assumption that the embedding space adequately captures traditional Chinese medical concepts may be wrong. Rather, concepts are probably mapped through the dominant -most frequent- culture conveyed by the documents used to train the models [143]. The risk is that we perceive that different cultural perspectives are now accessible while we are just accessing culturally-specific information filtered through the dominant cultural framework of the training data.

At the individual scale, individuals may have easier access to information previously limited by language barriers or by the need to combine too many tools or steps —multiple searches in multiple languages plus translation. But individuals access decontextualised information and may not have the cultural background needed to interpret it appropriately.

At the community scale, educational/research/etc. communities could develop more inclusive and less language/disciplinary/cultural silo-based curriculae/research/etc. However, systems may prioritize some views/cultures.

At the societal scale, there is the opportunity to preserve minority knowledge/culture/languages for not much additional costs if we avoid digital colonization and a dominant culture being represented.

Knowledge Authority vs Knowledge Diversity. At the production stage, information authority is established through editorial workflows which consists of a sequence of tasks and responsibilities in the creation, editing, and publication of content. Information authority can also be given through the official status that a government assigns to information, e.g., through legal force, the credibility of authorship, recognition by the domain or community, topicality, etc. Search engines reinforce this by giving more weight to some sources, e.g. through PageRank-based ranking, topicality consideration, etc. The IR community has developed many algorithms and techniques that prevail for ordering the retrieved results. On top of that, the engine financing model has led to additional ranking strategies, making the ranking non-transparent to users in the end. In RAG/GAR systems, the risk is that things are getting worse, there are no safeguards on how sources are prioritized.

RAG systems potentially democratize knowledge by treating all embedded documents as equally represented based on their semantic content, rather than institutional or other authority; although, not all languages/cultures are equally represented in embedding spaces. And some knowledge communities do not respond positively to digitization, specially minority communities.

But the retrieval part of RAG/GAR systems plays a crucial role in the drift of the generation part. And it should be more transparent, specially on how sources are selected and how many are used. A recent study analyzes [10] how ChatGPT sorts through the vast array of sources available on the Internet on the news summary task (here "What happened yesterday in Gaza?" in French from France on ChatGPT4.1). According to the report, "Considering only publications from the last 24 hours, Google News offers between 25 and 60 publications from the news media every day (45 on average), while ChatGPT offers between 5 and 10 (7.8 on average)—about six times less." The underlying mechanisms are black boxes. Either it should be more transparent or driven more by the user rather than keeping an automatic driver of the system. The RAG/GAR principle creates new opportunities for knowledge diversity but raises critical questions about validation at the different stages.

In many domains the "truth" or maybe rather the authoritative information may depend on the culture/location; sometimes for the good, sometimes for the bad.

At the individual scale, users face the challenge of evaluating unfamiliar knowledge systems without traditional authority markers. Relying on the system only could be even more risky.

At the community scale, institutions and communities should decide whether to maintain existing authority structures or develop new frameworks that recognize new and maybe more diverse forms of authorities.

At the societal scale, the risk can be of either enforcing dominant cultural authority standards globally while it could be useful to create systems where all knowledge claims are treated as equally valid.

4.4.8 Conclusion

Critical deliberations on the societal implications of technology can help us to free ourselves from technodeterministic thinking and encourage the community to actively assert their agency to shape technologies to affect desired positive social outcomes. The proposed shift from “RAG” to “GAR” is our attempt to refocus the discussion on our society’s information access needs rather than on the new technology itself. This needs to be accompanied by critical conversations within the IR community to explicate and reimagine our sociotechnical imaginaries [115], deliberate on the values we want to encode in our design (*e.g.*, democratic and emancipatory [166]), and view these platforms from diverse perspectives (*e.g.*, as critical

[10] `https://larevuedesmedias.ina.fr/chatgpt-gaza-google-actualites-information-sources`

infrastructure that should be open[11] and democratically governed). These conversations must not be restricted only to IR researchers and practitioners, but must encourage cross-disciplinary participation by scholars and experts from outside of IR.

This write-up is only the first step. We need to follow up with other concrete actions geared towards transforming not just what the IR community works on, but how it approaches the questions and with what goals. We must reflect on how the arguments presented in this section intersects with the conversations around other proposals, such as the #FreeWebSearch charter[12]. And we should be clear-eyed about the social, political, and cultural context in which this work is embedded and the incentive structures that continues to shape our work.

[11] https://freewebsearch.org/

[12] https://freewebsearch.org/en/charter/

4.5 An Unexamined RAG Is Not Worth Interrogating

Niklas Deckers (University of Kassel and hessian.AI, DE)
Laura Dietz (University of New Hampshire, USA)
Maik Fröbe (Friedrich-Schiller-Universität Jena, DE)
Wojciech Kusa (NASK National Research Institute, PL)
Mark Sanderson (RMIT University, AU)

4.5.1 Motivation

Modern tool chains make it relatively easy to quickly develop retrieval-augmented generation (RAG) systems. Hence, many RAG systems exist, for different applications, different corpora, and in different configurations. However, it is difficult to know which of these RAG systems is best, and for which contexts each is most appropriate. Therefore, we aim to answer two questions:

1. When RAG works, how can we be certain that it works?
2. When RAG fails, how can we find this out?

While there is extensive work on the evaluation of RAG, we believe that RAG as a technology is so new, it's still unclear what it is that we need to evaluate in a RAG system. We remain unconvinced that even state-of-the-art evaluation paradigms [54, 138, 194] have, as yet, demonstrated that they can measure all that we need to understand in RAG.

The paradigm shift that RAG presents to information access technology requires a corresponding shift in evaluation procedures. We have seen such shifts in the past: the Cranfield methodology for retrieval evaluation was adjusted to scale to larger document collections [173]. With RAG, we find ourselves in a similar situation that requires novel evaluation. Table 2 provides a comparison of aspects that we hypothesize might change substantially between information retrieval (IR) and RAG evaluation for online and offline experiments.

For online experiments, explicit click feedback in IR was exploited for learning at scale. With the reduction in clicks now found in RAG systems, implicit feedback might be more difficult to incorporate. As users express themselves in natural language, RAG systems can be personalized more easily as they can be explicitly prompted to follow users' instructions. For offline experiments, traditional retrieval evaluation had the advantage that the document collection was stable and the relevance judgments were per document so that benchmarks could be re-used to evaluate new systems over a long period. For RAG evaluations, it might be that evaluation benchmarks are more difficult to re-use as new RAG systems might produce unseen responses—even for the same query and collection. These unseen responses are possibly of equal quality, but utilize different language, different order, and different sources. This renders purely manual evaluations impractical. While it is still possible to apply manual evaluations to individual system responses (frozen in time), the costs are generally considered too high, given that resulting research is not reproducible and the test collections cannot be reused for developing the next generation of systems.

The rapid development of new innovations on the basis of LLMs and RAG must be met with novel approaches for measuring quality. Although prompt-based LLM-as-a-judge approaches are popular, they are vulnerable to issues such as LLM narcissism, circularity, and benchmark memorization, which lead to measurement errors [43].

Table 2 Asymmetry of retrieval and RAG evaluation across Offline and Online experiments.

	IR	RAG
(a) Offline Experiments: Explicit feedback at low quantity of high quality		
Longevity of Benchmarks	High	Low
Feedback Granularity	Per Document	Per System
(b) Online Experiments: Implicit feedback at high quantity of low quality		
Learning from Feedback	Easy	Difficult
Personalization	Difficult (Implicit)	Easy (Explicit)
(c) Both (Online and Offline)		
Evaluation Methodology	Established	Exploration
Feedback Modalities	Few	Many

We briefly detail current RAG evaluation approaches and outline research challenges. Finally, to reverse the tempting trend of removing human feedback from the evaluation setup, we suggest development of open source tooling that supports exploratory examination of RAG systems. The hope is that by supporting the rapid identification of common failure modes, we aid the subsequent development of improved evaluation paradigms to measure the progress of innovation.

4.5.2 Research Gap

In 2023, there was a paradigm shift in IR. Before the advent of LLMs, it was difficult to build systems that summarized retrieved content into fluent language or supported conversational clarifications from users. Nowadays, such challenges have promising solutions. As a result, peoples' expectations of retrieval have grown. It is not sufficient anymore to merely obtain information that contains topically relevant text among many non-relevant sections. Instead, searchers expect that retrieved information is concise and on-point, while relating the content to the user's task, and potentially, their context.

4.5.2.1 State-Of-The-Art in RAG Evaluation

Distinct from IR, RAG tends to produce a different response every time the system is run—even for the same query and corpus. Initial attempts of addressing the variability of system responses with ROUGE/METEOR [18, 103], BERTScore [193], etc. have largely been abandoned, due to their inability to generalize to the variety of responses.

As a result, many researchers have abandoned the idea of grounding evaluation with humans, and moved towards fully automatic evaluation with the aid of LLM judges [54, 194]. Hybrid evaluation solutions have been proposed [138]. However, because of low cost and versatility, the adoption of fully automatic evaluation is popular despite the many already stated issues. Evaluation systems are even adopting an LLM judge idea to generate training data, which risks circularity, overfitting, and self-training collapse [43].

One compromise that lets humans contribute to a semi-automatic evaluation paradigm is an evaluation that focuses on a set of information nuggets or grading rubrics. The paradigm gives a human judge the ability to specify which pieces of information must, should, or could appear in the response of an excellent system. In this way the judge defines the required content. Such pyramid-based evaluation systems [104] have long been discussed, but adoption was largely hindered by the cumbersome manual process of aligning nuggets to text, requiring more work than a straight-up relevance judgment. By leveraging LLMs for the linguistic match between each nugget and corresponding passages in the system response, this hurdle has been overcome [55, 132, 139].

The main advantage of RAG systems over purely generative systems is the grounding of information in high-quality documents, increasing the trust that information is factually correct. One important part of RAG evaluation also verifies the faithfulness of citations.

Evaluation systems such as ARGUE and AutoArgue [114, 174] combine multiple metrics, such as nugget coverage, verification of citation support, and relevance of cited documents. While fulfilling these criteria is necessary for a high-quality system response, researchers realize that this list is not complete.

4.5.3 Proposed Research Directions

In order to develop the next generation of evaluation methods, we need to first understand the scope of open issues that state-of-the-art RAG systems should be measured against. A number of reviews of RAG evaluation have organized the issues differently. Three reviews/reflections [6, 7, 166] considered evaluation from the same two perspectives: evaluation of RAG systems and how generative AI can supplement the evaluation process itself including the possibility of creating digital twins, simulating a range of human interactions with information access systems. A more recent review identified four areas: evaluation of benchmark datasets, of indexing, of the retriever, and finally evaluation of the generator [21]. Recently, novel evaluation methodologies, such as red teaming [9], have been applied to RAG systems.

During the seminar, we examined what we saw as overlooked aspects of evaluation: how to encourage the users of RAG systems to evaluate the content that they are shown and how to support evaluators of RAG systems to interrogate the correctness of answers.

4.5.3.1 Encouraging Users to Be Evaluators

LLMs are trained to produce confident and assertive language that is designed to convince humans—even when the level of confidence does not match the content quality. This property of LLMs can be exploited by a range of stakeholders for advertisement, opinionated debate, job search and hiring, and management-related messaging. Therefore, users might not realize the quality (or lack thereof) of the provided information. Means of encouraging the users of RAG systems to more fully interrogate answers is critical. The current approach used by many RAG systems of displaying a warning that "AI responses may include mistakes" is not a sufficient solution. As RAG systems improve, we need to detect whenever quality criteria have been "addressed" (e.g. fluency), and whenever new criteria arise (e.g. plausibility vs correctness) that determine the difference between best and mediocre systems.

Given the tendency of people to over-rely on LLMs [59], we need to consider what might motivate users to carefully examine the validity of RAG system responses. This would enable us to design gamified evaluation methods [197] and vigilance tests [33]. Leveraging interaction data. With the rise of AI overviews emanating from RAG systems, there has been a measured drop in click data [29], which have been an essential signal for online evaluation. Hence, there is a pressing need for collecting new forms of user interaction data for new approaches towards online evaluation. An example of such a novel problem is the emergence of sycophancy in the output of generative systems [32].

While this categorization of RAG systems is a good start, the novelty of RAG is such that there may be uses and impacts of this technology that are as yet unknown. Charting such a landscape requires a different approach. There are research methodologies that can help establish such a landscape. Qualitative approaches that engage with users and their data in an exploratory manner can be used to identify potential research problems that can be examined afterwards.

4.5.3.2 Supporting RAG System Evaluators

In the past, measuring quality was largely a question of costs: how much annotator time to purchase? Do annotators require specialized knowledge (e.g., biomedical experts)? Are crowdworkers sufficient? How many topics and responses of the assessment pool to judge?

A general issue that has always affected traditional IR evaluation is that human judges can only measure "recall" with respect to information provided in an assessment pool. Relying on the world knowledge of judges when assessing the relevance of non-contextualized facts, often led to non-ideal results [42]. In a world where RAG systems are all providing—at a minimum—superficially correct appearing responses, even a careful assessor might not be able to notice omitted or false information. Even when citations are verified for faithfulness, the citation of documents with incorrect or unhelpful information yields RAG responses of low quality.

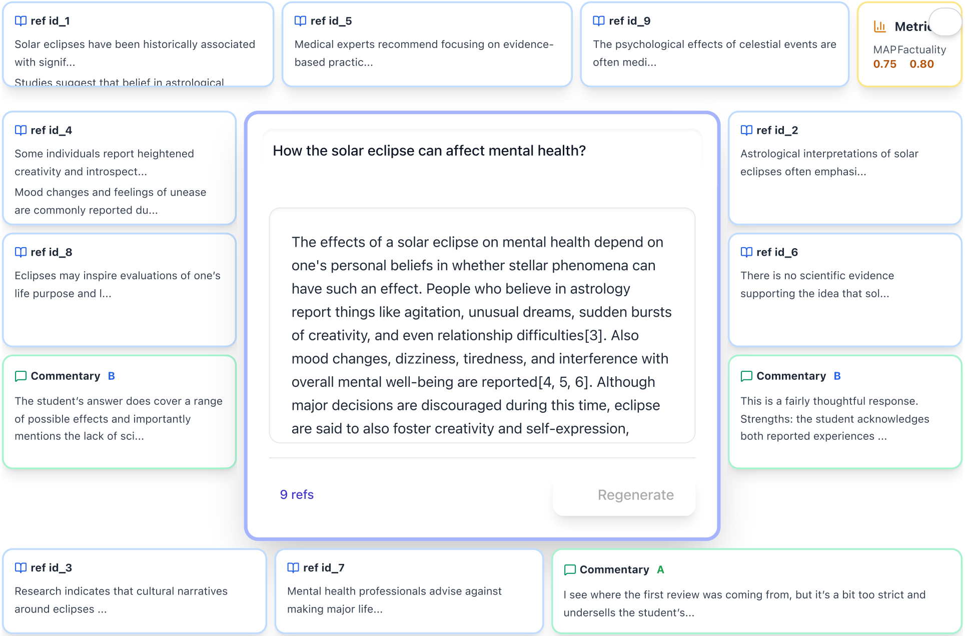
It will be difficult (or maybe impossible) to measure the quality with human judges or LLMs alike. Moreover, any evaluation that is purely based on LLM judges, will likely be susceptible to the same issues as the response-generating LLM itself. We suggest to invest more research into how to reveal subtle quality differences between RAG/IR systems for such situations without obvious solutions. We hypothesize that the solution will likely involve a collaboration between humans and LLMs. A complementary approach is task-based evaluation, in which humans attempt difficult tasks with RAG assistance, revealing system quality through its impact on real problem-solving.

We need more research to identify the properties via which quality differences manifest and design experimental setups to reveal such differences. This will require to think which information human judges and LLM judges need to be presented with to focus on judging such properties. It also will require to define human-LLM co-working processes that effectively compensate for each other's failure modes.

4.5.4 Talmud-Style Interfaces to Discuss RAG Response Quality

As issues get addressed and RAG systems evolve, we need an easy way to identify and prioritize the most urgent failure modes of RAG systems. We assume that human system developers and judges need support to manually inspect RAG system responses and discuss the pros, cons, and failures in a systematic fashion. This support may come through the way that this information is presented. During the workshop, we got inspiration from religious and ethical discourses that have experience of keeping discussions on important topics ongoing for hundreds of years. Therefore, those ideas inspired our discussions.

As an approach, we draw on the analogy of how the Talmud is used for discussion over hundreds of years [91] and apply it to RAG quality assessment interfaces. The Talmud is composed of two parts: The *Mishnah* contains the laws that are derived from the Torah, and the *Gemara* provides interpretation and discussion of those laws. These are presented in an integrated way, with the text of the Mishna shown in the center and the discussion of the Gemara presented as annotations in the margins. The interpretations given here are not considered absolute; they reflect different perspectives.



Solar eclipses have been historically associated with signif...

Studies suggest that belief in astrological

Medical experts recommend focusing on evidence-based practic...

The psychological effects of celestial events are often medi...

Some individuals report heightened creativity and introspect...

Mood changes and feelings of unease are commonly reported du...

Eclipses may inspire evaluations of one's life purpose and l...

The student's answer does cover a range of possible effects and importantly mentions the lack of sci...

How the solar eclipse can affect mental health?

The effects of a solar eclipse on mental health depend on one's personal beliefs in whether stellar phenomena can have such an effect. People who believe in astrology report things like agitation, unusual dreams, sudden bursts of creativity, and even relationship difficulties[3]. Also mood changes, dizziness, tiredness, and interference with overall mental well-being are reported[4, 5, 6]. Although major decisions are discouraged during this time, eclipse are said to also foster creativity and self-expression,

Astrological interpretations of solar eclipses often emphasi...

There is no scientific evidence supporting the idea that sol...

This is a fairly thoughtful response. Strengths: the student acknowledges both reported experiences ...

Research indicates that cultural narratives around eclipses ...

Mental health professionals advise against making major life...

I see where the first review was coming from, but it's a bit too strict and undersells the student's...

Figure 9 Overview of the Talmud-inspired user interface for RAG quality discussions. At the center of the interface is a user query accompanied by a single model-generated response. Surrounding it are the key components of the quality discussion: human and LLM discussion of pros/cons, the original snippets, and automated evaluation metric scores.

Following this structure, we envision a RAG annotation interface in which the central part contains a user query and the RAG system response. In the margins, retrieved snippets provide the evidential basis for the generation, supporting or contextualizing its content. Beyond the snippets, a further layer of commentary—resembling the *Gemara*—may include discussion among human experts aided by LLM-as-a-judge assessments, offering interpretative assessments of the answer's quality. Additionally, aggregated evaluation metrics can be presented as a meta-view of system performance. Figure 9 illustrates the vision of such a Talmud-inspired design.

An alternative configuration of the interface would invert the layering. Instead of focusing on the RAG response, we could place a cited document in the center (analogous to the *Mishnah*), with multiple system generations arranged as form of discussion (analogous to the *Gemara*).

In the Talmudic tradition, such higher-order reflections are captured in later super-commentaries (e.g., by the *Rishonim* and *Acharonim*), which offer critical perspectives on both the Mishnah and the Gemara. In a similar way, the meta-assessment layer in our envisioned system functions as a set of super-commentaries, guiding assessors in contextualizing, comparing, and ultimately judging the value of different generations.

Building on this design analogy, the proposed system can also be situated within established IR evaluation traditions. Our "Talmud-inspired" interface reframes evaluation by highlighting both the core unit of analysis (query and system response) and the multilayered interpretative context (retrieved snippets, intermediate-steps, commentaries, and meta-assessments). Such a system could be used by several user groups:

- Assessors who judge the quality of system outcomes across multiple models are benefiting from the structured presentation of supporting evidence, commentaries, and aggregated metrics. LLMs can help to maintain consistency among commentary by identifying when earlier comments apply to new responses.
- Evaluation developers, who study the reliability of human and machine assessors in layered evaluation settings, in addition to overseeing the assessment process.
- Professional searchers in domain-specific applications (e.g., legal, medical, scientific retrieval), who require both direct answers and transparent supporting evidence to make informed decisions.
- Educators and learners (e.g., medical students, legal clerks) who study the discourse in complex systems may benefit from an integrated presentation of both authoritative sources (Mishnah), multiple perspectives and interpretations (Gemara), as well as higher-level critiques (super-commentaries).
- Users who prefer a presentation of the structured dialogical space [15] over a linear ranking by relevance [81, 190].

4.5.5 Broader Impact

There are two main areas of impact for the research agenda that is described: researchers working on RAG systems and users of RAG technology.

Current methods for assessing these systems are still much in their infancy. Methodologies such as the Cranfield Paradigm—which has served the information retrieval community exceptionally well for over six decades—appears to be reaching a distinct limit: just simply focusing on the relevance of objects is no longer sufficient to understand the value of the output of a generative information access system, such as RAG. We are calling for further research on how to better assess the quality of the RAG systems. This includes systems being developed currently as well as the systems that will be designed in the future.

The potential impacts of the research agenda we detail here are broad. Features of RAG technology are becoming increasingly visible with search engines, such as Google or Bing, which are starting to roll out the AI summaries in a large number of search results [29]. Many desktop applications recently incorporated their own “AI feature”. This, however, is only the tip of the iceberg. RAG is proving to be an extremely popular technology. It is being deployed by a wide range of organizations for the management, search, summarization, and manipulation of documents. While well funded companies such as Google can afford to spend substantial sums of money to ensure their RAG systems are accurate, all organizations using RAG will also want to ensure that the systems they manage are operating successfully. Our proposed research agenda will benefit all stakeholders of deployed RAG systems.

5 Answers to "Will RAG replace ranked search for end users?"

Vote	**Reasoning** (Colors: negative, positive)
No	*No reason given*
No	For navigational queries, RAG is plain stupid (Google summaries sometimes suggest ¨doing a Google search" for such queries, which is hilarious). RAG without links is no search at all, and mostly useless, unless you already know the answer (known item ¨search").
No	RAG systems are usually not transparent about their retrieval results before the generation step. As long as that doesn't change, fact checking will always be done post-hoc in an ad-hoc search engine.
No	It would be incredibly problematic if we can only imagine futures where all information access is intermediated by LLMs that concentrates absolute power over what and how information is presented, that serves as a mechanism for mass manipulation of public opinion, in the hands of the platform owners. While there are significant societal concerns with any use of LLMs, the notion that they completely replace all forms of information access modalities is particularly dangerous.
No	I assume that end users (depending on the task) would still like the option to select the sources they value more.
No	I think there will still be a need for keyword-to-ranked-documents type of systems, e.g., for navigational queries.
No	No, because for some scenarios, expressing the search intent verbally with sufficient precision is more difficult than quickly skimming search snippets, and/or the generated text is structured less intuitively than a 10-blue-links format.
No	Rather than replace it, I believe that they could complement each other. * Depending on the intent, a user might want to get a large list of results (e.g., reviews, lawsuits) and remain in control regarding which results to access and assess themselves relevance and trustworthiness. Especially as RAG can exhibit hallucinations. * From a sustainability perspective, ranked search typically requires less resources than RAG (e.g., computational power, electricity). Finding a compromise can help to optimize resource usage.
No	Certainly in some search applications, it definitely will. We're already seeing this with web search. I think there will still be cases where traditional search remains: e.g., specialist search applications (e.g., patient search, scientific literature search, etc.), navigational search (e.g., product search, website search, etc.). But these will be in the vast minority.
No	For question-answering types of intents, RAG could replace ranked search, but for decision-making intents, no.
No	RAG matches many targeted information needs well. However, many others benefit from having multiple results; be it for diversity, for easy verification, or otherwise.
No	It will replace it to a good extent (for many small or "superficial" searches), but not fully (for in-depth research or search for original information or content).
No	It is unlikely that RAG replaces search for all domains (e.g., legal domain). I, however, believe the future of search is conversational, where system utterances are not just RAG results, but also high precision search results.

Vote	**Reasoning** (Colors: negative, positive)
No	As RAG gets better, more and more information needs will be answered by a generated answer rather than a traditional ranked list of results. Yet, a number of information seeking tasks are intrinsically required to retrieve documents, not answers—and thus ranked lists are a good means to satisfy these requests. For example for navigational and transactional queries, a ranked list seems more efficient than a RAG output. Currently, also queries that require a high degree of comparison and exploration are better answered via RAG—however this is most likely to change.
No	Predicting is hard, especially if it's about the future... :) But: Did the calculator replace manual arithmetic-operations for all end users? Did the typewriter or the computer replace hand writing? RAG may reduce the use of ranked search, but I doubt it will fully replace the exposure of rankings to all end users.
No	It will replace ranked search for users prioritizing convenience. For those with other priorities (e.g., being able to directly access primary sources, reducing the ecological impact of their searches), ranked search will not be replaced by RAG.
Yes	Most end users will become to lazy to check the underlying info. However, professional users (e.g. in the medical domain) will remain skeptical and ask for transparency.
Yes	In many information-seeking tasks and scenarios, a RAG system will be the users' preferred means of finding information. That doesn't mean that ranking will be gone completely, but the major preference would be towards RAG-based systems.
Yes	For many search settings, yes. We're already seeing it with web search. There will likely always be applications where users still get ranked search result lists, though (e.g., navigational, patent search, product search, etc.).
Yes	Because RAG basically entails ranked search. If quality problems are solved, it is just another UI to ranked search which users have to get used to.
Yes	You force me to say Yes or No, so I pick yes, but honestly the answer is I think mostly. Of course most RAG systems have ranked retrieval as a core part of their infrastructure, so is that actually a replacement?
Yes	I wish there had been an "It depends" or "not sure" response above :) I feel like this is where things are heading – technology is evolving quickly and people want to do things quickly and efficiently.
Yes	Lacking technological expertise this is only a guess or rather a wish because RAG represents from an ethical/pedagogical perspective a specific order of knowledge and a culture of "knowing" that I find tempting.
Yes	It may just become a subcomponent of larger chatbot pipelines... only few users will want to take the pain of searching themselves and corroborating results with LLM answers.
Yes	As an augmented feature with ranked search, RAG has a huge potential to re-think/change rank search for end users.

Vote	**Reasoning** (Colors: negative, positive)
Yes	I rarely use web search anymore, and when I do, I am struck by how degraded the experience has become. The rise of AI-generated "slop" has made search results nearly unusable, but the deeper issue is that search engines have failed (or chosen not) to adapt their spam detection to this new reality. Because of this, the industry seems to be drifting toward retrieval-augmented generation (RAG) as a replacement for search. I am uneasy about this direction, particularly given the higher cost and energy demands of RAG compared to traditional retrieval. My original hope for RAG was that it would serve primarily as a retrieval system with a light layer of contextual interpretation and summarization. Instead, many implementations now generate an answer first and then search for supporting sources afterward — a reversal that raises obvious concerns about reliability and trustworthiness. — this summary was made readable by GPT5
Yes	*No reason given*
Yes	*No reason given*
Yes	*No reason given*
Yes	*No reason given*

6 Recommended Reading List

These publications were recommended by the seminar participants via the pre-seminar survey.

- Negar Arabzadeh and Charles L. A. Clarke. 2025. Benchmarking LLM-Based Relevance Judgment Methods. In *Proceedings of SIGIR 2025*, pages 3194–3204. `https://doi.org/10.1145/3726302.3730305`
- Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-Based Language Models and Applications. In *Proceedings of ACL 2025*, pages 41–46. `https://doi.org/10.18653/V1/2023.ACL-TUTORIALS.6`
- Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, Judges, and Assistants: Towards Understanding the Interplay of LLMs in Information Retrieval Evaluation. In *Proceedings of SIGIR 2025*, pages 3865–3875. `https://doi.org/10.1145/3726302.3730348`
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of ICML*, pages 2206–2240. `https://proceedings.mlr.press/v162/borgeaud22a.html`
- Alissa Centivany. 2024. Mining, Scraping, Training, Generating: Copyright Implications of Generative AI. In *Proceedings of the Association for Information Science and Technology*, 61(1):68–79. `https://doi.org/10.1002/pra2.1009`

- Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting Conversational Information Seeking Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. In *Proceedings of CHIIR 2024*, pages 401–405. `https://doi.org/10.1145/3627508.3638309`
- Charles Clarke and Laura Dietz. 2025. LLM-Based Relevance Assessment Still Can't Replace Human Relevance Assessment. In *Proceedings of EVIA 2025*. `https://doi.org/10.20736/0002002105`
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of SIGIR 2024*, pages 719–729. `https://doi.org/10.1145/3626772.3657834`
- Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of ICTIR 2025*, pages 218–229. `https://doi.org/10.1145/3731120.3744588`
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of KDD 2024*, pages 6491–6501. `https://doi.org/10.1145/3637528.3671470`
- Naghmeh Farzi and Laura Dietz. 2025. Criteria-Based LLM Relevance Judgments. In *Proceedings of ICTIR 2025*, pages 254–263. `https://doi.org/10.1145/3731120.3744591`
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997*. `https://doi.org/10.48550/ARXIV.2312.10997`
- Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of SIGIR 2025*, pages 1916–1929. `https://doi.org/10.1145/3626772.3657849`
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of ICML 2020*, pages 3929–3938. `https://dl.acm.org/doi/abs/10.5555/3524938.3525306`
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55. `https://doi.org/10.1145/3703155`
- Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of EMNLP 2023*, pages 7969–7992. `https://doi.org/10.18653/V1/2023.EMNLP-MAIN.495`

- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP 2020*, pages 6769–6781. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550
- Carol Collier Kuhlthau. 1993. A Principle of Uncertainty for Information Seeking.
- J. Documentation, 49(4):339–355. https://doi.org/10.1108/EB026918
- Weronika Lajewska and Krisztian Balog. 2025. GINGER: Grounded Information Nugget-Based Generation of Responses. In *Proceedings of SIGIR 2025*, pages 2723–2727. https://doi.org/10.1145/3726302.3730166
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of NeurIPS 2020*, pages 9459-9474. https://dl.acm.org/doi/abs/10.5555/3495724.3496517
- Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. 2024. Sociotechnical Implications of Generative Artificial Intelligence for Information Access. *arXiv:2405.11612*. https://doi.org/10.48550/ARXIV.2405.11612
- Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flávio du Pin Calmon, Margaret Mitchell, Michael D. Ekstrand, Reuben Binns, and Solon Barocas. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. *arXiv:2506.14652*. https://doi.org/10.48550/ARXIV.2506.14652
- Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-Response. In *Proceedings of ACL 2011*, pages 1415–1424. https://aclanthology.org/P11-1142/
- Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's High Time: A Survey of Temporal Information Retrieval and Question Answering. *arXiv:2505.20243*. https://doi.org/10.48550/ARXIV.2505.20243
- Martin Potthast, Matthias Hagen, and Benno Stein. 2020. The Dilemma of the Direct Answer. In *ACM SIGIR Forum*, 54(1):14:1–14:12. https://doi.org/10.1145/3451964.3451978
- Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial Nugget Evaluation Results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. *arXiv:2411.09607*. https://doi.org/10.48550/ARXIV.2411.09607
- Mandeep Rathee, V. Venktesh, Sean MacAvaney, and Avishek Anand. 2025. Test-Time Corpus Feedback: From Retrieval to RAG. *arXiv:2508.15437*. https://doi.org/10.48550/arXiv.2508.15437
- Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of SIGIR 2024*, pages 2395–2400. https://doi.org/10.1145/3626772.3657957
- Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *Proceedings of SIGIR 2022*, pages 2825–2837. https://doi.org/10.1145/3477495.3531766
- Tobias Schreieder, Tim Schopf, and Michael Färber. 2025. Attribution, Citation, and Quotation: A Survey of Evidence-Based Text Generation with Large Language Models. *arXiv:2508.15396*. https://doi.org/10.48550/arXiv.2508.15396

- Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of CHIIR 2022*, pages 221–232. https://doi.org/10.1145/3498366.3505816
- Chirag Shah and Emily M. Bender. 2024. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? In *ACM Trans. Web*, 18(3):33:1–33:24. https://doi.org/10.1145/3649468
- Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of CHI 2024*, pages 1–17. https://dl.acm.org/doi/10.1145/3613904.3642459
- Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong, and Yiqun Liu. 2025a. Dynamic and Parametric Retrieval-Augmented Generation. In *Proceedings of SIGIR 2025*, pages 4118–4121. https://doi.org/10.1145/3726302.3731692
- Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models Can Accurately Predict Searcher Preferences. In *Proceedings of SIGIR 2024*, pages 1930–1940. https://dl.acm.org/doi/10.1145/3626772.3657707
- Johanne R. Trippas, J. Shane Culpepper, 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, 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 P. De Vries, Ryen W. White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Justin Zobel, Shengyao Zhuang, and Guido Zuccon. 2025. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *ACM SIGIR Forum*, 59(1). https://www.johannetrippas.com/papers/trippas2025swirl.pdf
- Venktesh Viswanathan, Mandeep Rathee, and Avishek Anand. 2025. Trust but Verify! A Survey on Verification Design for Test-Time Scaling. *arXiv:2508.16665*. https://doi.org/10.48550/arXiv.2508.16665
- Ellen M. Voorhees. 2002. The Evaluation of Question Answering Systems: Lessons Learned from the TREC QA Track. In *Proceedings of LREC 2002*, page 6. http://www.lrec-conf.org/proceedings/lrec2002/pdf/ws7.pdf
- Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness Is Not Faithfulness in RAG Attributions. In *Proceedings of ICTIR 2025*, pages 22–32. https://dl.acm.org/doi/10.1145/3731120.3744592
- Ryen W. White and Chirag Shah. 2025. Information Access in the Era of Generative AI. *Springer*. https://doi.org/10.1007/978-3-031-73147-1
- Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. 2024. Retrieval-Augmented Generation for Natural Language Processing: A Survey. *arXiv:2407.13193*. https://doi.org/10.48550/ARXIV.2407.13193
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR 2023*. https://openreview.net/forum?id=WE_vluYUL-X

7 Acknowledgments

The seminar organizers would like to thank all participants and speakers of invited talks for their active contributions. We also thank the staff of Schloss Dagstuhl for providing a great venue for a successful seminar. The organizers were in part supported by the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014), the BMFTR project CORAL under FKZ 011S24077-B, and the BMFTR project DIALOKIA under FKZ 01IS24084A-B. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

Participants

- Qinqyao Ai
Tsinghua University, CN
- Mohammad Aliannejadi
University of Amsterdam, NL
- Liesbeth Allein
KU Leuven, BE
- Sophia Althammer
Cohere, DE
- Avishek Anand
Delft University, NL
- Nolwenn Bernard
TH Köln, DE
- Niklas Deckers
University of Kassel and hessian.AI, DE
- Gianluca Demartini
The University of Queensland – Brisbane, AU
- Laura Dietz
University of New Hampshire, US
- Carsten Eickhoff
Universität Tübingen, DE
- Nicola Ferro
University of Padova, IT
- Maik Fröbe
Friedrich-Schiller-Universität Jena, DE
- Norbert Fuhr
Universität Duisburg-Essen, DE
- Marcel Gohsen
Bauhaus-Universität Weimar, DE
- Michael Granitzer
University of Passau, DE
- Faegheh Hasibi
Radboud Universtiy, NL
- Sebastian Heineking
Universität Leipzig, DE
- Djoerd Hiemstra
Radboud University, NL
- Adam Jatowt
Universität Innsbruck, AT
- Abhinav Joshi
Indian Institute of Technology Kanpur, IN
- Johannes Kiesel
GESIS – Leibniz Institute for the Social Sciences, DE
- Wojciech Kusa
NASK National Research Institute, PL
- Sean MacAvaney
University of Glasgow, UK
- Bhaskar Mitra
Independent Researcher, Tiohtià:ke/Montréal, CA
- Josiane Mothe
INSPE, Université de Toulouse, UT2J, UMR5505 CNRS IRIT, FR
- Smaranda Muresan
Barnard College, Columbia University, US
- Jian-Yun Nie
University of Montreal, CA
- Heather O'Brien
iSchool, University of British Columbia, CA
- Birte Platow
TU Dresden, ScaDS.AI, DE
- Martin Potthast
Universität Kassel, hessian.AI, ScaDS.AI, DE
- Mark Sanderson
RMIT University – Melbourne, AU
- Harrisen Scells
Universität Tübingen, DE
- Alan Smeaton
Dublin City University, IE
- Damiano Spina
RMIT University – Melbourne, AU
- Benno Stein
Bauhaus-Universität Weimar, DE
- Johanne Trippas
RMIT University – Melbourne, AU
- Stefan Voigt
Open Search Foundation, DE
- Arjen P. de Vries
Radboud Universtiy, NL
- Guido Zuccon
Google Research Australia and The University of Queensland, AU

References

1 Abdelrahman Abdallah, Mahmoud Abdalla, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankarena: A unified platform for evaluating retrieval, reranking and rag with human and llm feedback, 2025.

2 Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. *arXiv preprint arXiv:2502.02464*, 2025.

3 Syed Rameel Ahmad. Enhancing multilingual information retrieval in mixed human resources environments: A rag model implementation for multicultural enterprise. *arXiv preprint arXiv:2401.01511*, 2024.

4 Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

5 Liqaa Habeb Al-Obaydi and Marcel Pikhart. Artificial intelligence addiction: exploring the emerging phenomenon of addiction in the ai age. *AI & SOCIETY*, pages 1–17, 2025.

6 Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. Generative information retrieval evaluation. In *Information access in the era of generative ai*, pages 135–159. Springer, 2024.

7 James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. Future of information retrieval research in the age of generative ai. *CoRR*, 2024.

8 Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):1–33, 06 2025.

9 Bang An, Shiyue Zhang, and Mark Dredze. RAG LLMs are not safer: A safety analysis of retrieval-augmented generation for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5444–5474, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

10 Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. Explainable information retrieval: A survey. *arXiv preprint arXiv:2211.02405*, 2022.

11 Anthropic. Claude 3.7 sonnet and claude code, February 2025. Accessed: 2025-09-25.

12 Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. Gar-meets-rag paradigm for zero-shot information retrieval. *CoRR*, abs/2310.20158, 2023.

13 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.

14 Imane El Atillah. Man ends his life after an ai chatbot 'encouraged' him to sacrifice himself to stop climate change, 2023.

15 Mikhail Bakhtin. *Problems of Dostoevsky's Poetics.* University of Minnesota Press, 1984.

16 Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, 2024.

17 Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. Reconsidering LLM uncertainty estimation methods in the wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29531–29556. Association for Computational Linguistics, July 2025.

18 Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

19 Marcia J. Bates. Where should the person stop and the information search interface start? *Inf. Process. Manag.*, 26(5):575–591, 1990.

20 Marshall Breeding. The future of library resource discovery. *Information Standards Quarterly*, 27(1):24–30, 2015.

21 Lorenz Brehme, Thomas Ströhle, and Ruth Breu. Can llms be trusted for evaluating rag systems? a survey of methods and datasets. In *Accepted for presentation at the IEEE Swiss Conference on Data Science (SDS25)*, 2025.

22 John Seely Brown and Paul Duguid. The social life of documents; introduction by esther dyson. *First monday*, 1996.

23 Matthew Burtell and Thomas Woodside. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

24 Jamie Callan. Distributed information retrieval. In *Advances in information retrieval: Recent research from the Center for Intelligent Information Retrieval*, pages 127–150. Springer, 2002.

25 Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13, 2023.

26 Lucia Casiraghi, Eugene Kim, and Noriko Hara. Tweeting on thin ice: Scientists in dialogic climate change communication with the public. *First Monday*, 2024.

27 Seung Woo Chae, Noriko Hara, Harshit Rakesh Shiroiya, Janice Chen, and Ellen Ogihara. Being vulnerable with viewers: Exploring how medical youtubers communicated about covid-19 with the public. *PloS one*, 19(12):e0313857, 2024.

28 Tyler A Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. *arXiv preprint arXiv:2410.17413*, 2024.

29 Athena Chapekis and Anna Lieb. Google users are less likely to click on links when an ai summary appears in the results. Technical report, Pew Research Center, 2025.

30 Catherine Chen, Jack Merullo, and Carsten Eickhoff. Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1401–1410, 2024.

31 Myra Cheng, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu. Dehumanizing machines: Mitigating anthropomorphic behaviors in text generation systems. *arXiv preprint arXiv:2502.14019*, 2025.

32 Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Social sycophancy: A broader anderstanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*, 2025.

33 Victoria L Claypoole, Daryn A Dever, Kody L Denues, and James L Szalma. The effects of event rate on a cognitive vigilance task. *Human factors*, 61(3):440–450, 2019.

34 Cyril Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. *Aslib-Cranfield Reseach Report, Cranfield*, 1962.

35 Merlise Clyde and Edward I George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

36 Simon Coghlan, Hui Xian Chia, Falk Scholer, and Damiano Spina. Control search rankings, control the world: what is a good search engine? *AI Ethics*, 5(4):4117–4133, 2025.

37 Daniel Cohen, Bhaskar Mitra, Oleg Lesota, Navid Rekabsaz, and Carsten Eickhoff. Not all relevance scores are equal: Efficient uncertainty and calibration modeling for deep retrieval models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 654–664, 2021.

38 Gemma Conroy. How chatgpt and other ai tools could disrupt scientific publishing. *Nature*, 622(7982):234–236, 2023.

39 Benoit Courty, Victor Schmidt, and et al. Sasha Luccioni. Codecarbon v2.4.1, 2024.

40 Julian De Freitas, Zeliha Oğuz-Uğuralp, and Ahmet Kaan-Uğuralp. Emotional manipulation by ai companions. Technical report, Harvard Business School Working Paper, 2025.

41 Robin De Mourat, Donato Ricci, and Bruno Latour. How does a format make a public? *Reassembling scholarly communications: Histories, infrastructures, and global politics of open access*, pages 103–12, 2020.

42 Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. Trec complex answer retrieval overview. In *TREC*, 2018.

43 Laura Dietz, Oleg Zendel, Peter Bailey, Charles LA Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. Principles and guidelines for the use of llm judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 218–229, 2025.

44 David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):45–70, 1995.

45 Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, 2024.

46 Martin Dugas, Philipp Neuhaus, Alexandra Meidt, Justin Doods, Michael Storck, Philipp Bruland, and Julian Varghese. Portal of medical data models: information infrastructure for medical research and healthcare. *Database*, 2016:bav121, 2016.

47 Maggie Harrison Dupré. Chatgpt is blowing up marriages as spouses use ai to attack their partners, 2025.

48 Maggie Harrison Dupré. People are being involuntarily committed, jailed after spiraling into “chatgpt psychosis”, 2025.

49 Yogesh K Dwivedi, Tegwen Malik, Laurie Hughes, and Mousa Ahmed Albashrawi. Scholarly discourse on genai's impact on academic publishing. *Journal of Computer Information Systems*, pages 1–16, 2024.

50 Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism Design for Large Language Models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, Singapore Singapore, May 2024. ACM.

51 Abul Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). *arXiv preprint arXiv:2505.02279*, 2025.

52 Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, et al. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*, 2024.

53 David Ellis. A behavioural approach to information retrieval system design. *J. Documentation*, 45(3):171–212, 1989.

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

55 Naghmeh Farzi and Laura Dietz. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 acm sigir international conference on theory of information retrieval*, pages 175–184, 2024.

56 Conor Feehly. Truth, romance and the divine: How ai chatbots may fuel psychotic thinking, 2025.

57 Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690. Association for Computational Linguistics, aug 2024.

58 Nicola Ferro and Carol Peters. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, volume 41. Springer, 2019.

59 Raymond Fok and Daniel S Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine*, 45(3):317–332, 2024.

60 Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In *International Conference on Theory and Practice of Digital Libraries*, pages 597–612. Springer, 2002.

61 Debasis Ganguly, Debarshi Kumar Sanyal, Prasenjit Majumder, Srijoni Majumdar, and Surupendu Gangopadhyay, editors. *FIRE '24: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation*. Association for Computing Machinery, 2024.

62 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

63 Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.

64 Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. *Foundations and Trends® in Information Retrieval*, 9(5):355–475, 2015.

65 Google. Supercharging search with generative ai, 2023. Accessed: 2025-09-25.

66 Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 478–479. ACM, 2004.

67 Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

68 Noriko Hara, Eugene Kim, Shohana Akter, and Kunihiro Miyazaki. Exploring the dynamics of interaction about generative artificial intelligence between experts and the public on social media. *Journal of Science Communication*, 24(1):A02, 2025.

69 Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as A daily assistant. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas, editors, *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 414:1–414:22. ACM, 2025.

70 Jeff Horwitz. A cognitively impaired new jersey man grew infatuated with "big sis billie," a facebook messenger chatbot with a young woman's persona. his fatal attraction puts a spotlight on meta's ai guidelines, which have let chatbots make things up and engage in 'sensual' banter with children"., 2025.

71 Allison Hosier and Laureen P Cantwell-Jurkovic. Ai and library and information science publishing: A survey of journal editors. *Library Trends*, 73(3):243–266, 2025.

72 Huang, S. and others. Values in the wild: Discovering and analyzing values in real-world language model interactions, April 2025. Accessed: 2025-09-25.

73 Niels-Henrik Höchstötter and Dirk Lewandowski. What users see: Structures in search engine results pages. *Information Sciences*, 179:1792–1812, 2009.

74 Peter Ingwersen and Kalervo Järvelin. *The Turn - Integration of Information Seeking and Retrieval in Context*, volume 18 of *The Kluwer International Series on Information Retrieval*. Kluwer, 2005.

75 Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, and Smaranda Muresan. Navigating the landscape of hint generation research: From the past to the future. *Transactions of the Association for Computational Linguistics*, 13:505–528, 2025.

76 Anubhav Jangra and Smaranda Muresan. Designing and evaluating hint generation systems for science education. In *submitted to CHIT 2026*, 2025.

77 Julie Jargon and Sam Kessler. A troubled man, his chatbot and a murder-suicide in old greenwich, 2025.

78 Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand, 2024. Association for Computational Linguistics.

79 Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, 2023.

80 Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv:2503.09516*, 2025.

81 Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.

82 Makoto P. Kato, Noriko Kando, Charles L. A. Clarke, and Yiqun Liu, editors. *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies.* National Institute of Informatics (NII), 2025.

83 Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

84 Fadhela Kerdjoudj and Olivier Curé. Evaluating uncertainty in textual document. In *URSW at ISWC*, 2015.

85 Dara Kerr. Musk's ai grok bot rants about 'white genocide' in south africa in unrelated chats, 2025.

86 Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions, 2025.

87 Johannes Kiesel, Çağrı Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Tomaž Erjavec, Matthias Hagen, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Harrisen Scells, Ines Zelch, Martin Potthast, and Benno Stein. Overview of Touché 2025: Argumentation Systems: Extended Abstract. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto, editors, *Advances in Information Retrieval*, volume 15576, pages 459–466. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

88 Miles Klee. He had a mental breakdown talking to chatgpt. then police killed him, 2025.

89 Miles Klee. People are losing loved ones to ai-fueled spiritual fantasies, 2025.

90 Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.

91 David C. Kraemer. *A History of the Talmud.* Cambridge University Press, 2019.

92 David R. Krathwohl. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218, 2002.

93 Carl Lagoze, Herbert Van de Sompel, Michael L. Nelson, and Simeon Warner. The open archives initiative protocol for metadata harvesting (OAI-PMH), version 2.0. Specification v2.0, Open Archives Initiative, June 2002. Released June 14, 2002.

94 Narendra Lahkar and Sanjib K Deka. Impact of query operators on web search engine results: An evaluative study. *Proceedings of the National Seminar on Informatics for Digital Information & Information Technology*, pages 141–149, 2004. Available via INFLIBNET.

95 Weronika Łajewska, Damiano Spina, Johanne Trippas, and Krisztian Balog. Explainability for transparent conversational information-seeking. In *Proceedings of the 47th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, page 1040–1050, 2024.

96 Brady D Land, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581, 2023.

97 Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. Can long-context language models subsume retrieval, rag, sql, and more? *CoRR*, abs/2406.13121, 2024.

98 Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024.

99 Stephen Ch Leung. The cognitive impacts of large language model interactions on problem solving and decision making using eeg analysis. *Frontiers in Computational Neuroscience*, 19:1556483, 2025.

100 Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac (ʻIkaʻaka) Nahuewai, Kari Noe, Danielle Olson, ʻŌiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI, 2020.

101 Jason Edward Lewis, Hēmi Whaanga, and Ceyda Yolgörmez. Abundant intelligences: placing ai within indigenous knowledge frameworks. *Ai & Society*, 40(4):2141–2157, 2025.

102 Andreas Liesenfeld and Mark Dingemanse. Rethinking open source generative ai: open-washing and the eu ai act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1774–1787, 2024.

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

104 Jimmy Lin and Dina Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 383–390, 2006.

105 Donna Lu. We tried out deepseek. it worked well, until we asked it about tiananmen square and taiwan, 2025.

106 Meng Lu, Catherine Chen, and Carsten Eickhoff. Cross-encoder rediscovers a semantic variant of bm25. *arXiv preprint arXiv:2502.04645*, 2025.

107 Jessica Lucas. What is ai-induced psychosis?, 2025.

108 Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, 24(253):1–15, 2023.

109 Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities, 2025.

110 Tamlin Magee. Mark sewards' ai misfire puts spotlight on bad chatbots, 2025.

111 Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online, August 2021. Association for Computational Linguistics.

112 Hannah R Marriott and Valentina Pitardi. One is the loneliest number… two can be as bad as one. the influence of ai friendship apps on users' well-being and addiction. *Psychology & marketing*, 41(1):86–101, 2024.

113 James Martin. *Managing the data base environment.* Prentice Hall PTR, 1983.

114 James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, et al. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1904–1915, 2024.

115 Bhaskar Mitra. Search and society: Reimagining information access for radical futures. *Information Retrieval Research*, 1(1):47–92, 2025.

116 Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. Sociotechnical implications of generative artificial intelligence for information access. In *Information Access in the Era of Generative AI*, pages 161–200. Springer, 2024.

117 Jonathan Morgan, Isaac Johnson, Michael Maggio, and Dario Taraborelli. Quantifying engagement with citations on wikipedia. In *Proceedings of the Web Conference 2020 (WWW '20)*, pages 1357–1368. ACM, 2020.

118 Josiane Mothe. Shaping the future of endangered and low-resource languages—our role in the age of llms: A keynote at ecir 2024. *SIGIR Forum*, 58(1):1–13, August 2024.

119 Jamshid Mozafari, Florian Gerhold, and Adam Jatowt. Wikihint: A human-annotated dataset for hint ranking and generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3821–3831, New York, NY, USA, 2025. Association for Computing Machinery.

120 Jamshid Mozafari, Anubhav Jangra, and Adam Jatowt. Triviahg: A dataset for automatic hint generation from factoid questions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2060–2070, New York, NY, USA, 2024. Association for Computing Machinery.

121 Jamshid Mozafari, Bhawna Piryani, Abdelrahman Abdallah, and Adam Jatowt. Hinteval: A comprehensive framework for hint generation and evaluation for questions. *CoRR*, abs/2502.00857, 2025.

122 Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for (llms) evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

123 Tori Noble and Kit Walsh. President trump's war on "woke ai" is a civil liberties nightmare, 2025.

124 OpenAI. How people are using chatgpt, 2025. Accessed: 2025-09-25.

125 OpenAI. Introducing chatgpt agent: bridging research and action, 2025. Accessed: 2025-09-25.

126 OpenAI. Introducing deep research, 2025. Accessed: 2025-09-25.

127 Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

128 Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15, 2024.

129 Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

130 Laura Perez-Beltrachini and Mirella Lapata. Uncertainty quantification in retrieval augmented question answering. *arXiv preprint arXiv:2502.18108*, 2025.

131 Peter L. T. Pirolli. *Information Foraging Theory*. Oxford University Press, May 2007.

132 Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–190, 2025.

133 Mark Raasveldt, Pedro Holanda, Tim Gubner, and Hannes Mühleisen. Fair benchmarking considered difficult: Common pitfalls in database performance testing. In *Proceedings of the Workshop on Testing Database Systems*. ACM, 2018.

134 Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*, 2025.

135 Cyrus Rashtchian and Da-Cheng Juan. Deeper insights into retrieval augmented generation: The role of sufficient context. `https://research.google/blog/deeper-insights-into-retrieval-augmented-generation-the-role-of-sufficient-context/`, 2025. Accessed: 2025-09-24.

136 Mandeep Rathee, V. Venktesh, Sean MacAvaney, and Avishek Anand. Test-Time Corpus Feedback: From Retrieval to RAG. *arXiv:2508.15437*, August 2025.

137 Haggai Roitman, Shai Erera, and Bar Weiner. Robust standard deviation estimation for query performance prediction. In *Proceedings of the acm sigir international conference on theory of information retrieval*, pages 245–248, 2017.

138 Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, 2024.

139 David P Sander and Laura Dietz. Exam: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In *DESIRES*, pages 136–146, 2021.

140 Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. Reduce, Reuse, Recycle: Green Information Retrieval Research. In *Proceedings of SIGIR*, pages 2825–2837, 2022.

141 Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Detecting Generated Native Ads in Conversational Search. In *Companion Proceedings of the ACM Web Conference 2024*, pages 722–725, Singapore Singapore, May 2024. ACM.

142 Chirag Shah and Emily M Bender. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Transactions on the Web*, 18(3):1–24, 2024.

143 Xinyang Shan, Yuanyuan Xu, Yining Wang, Yin-Shan Lin, and Yunshi Bao. Cross-cultural implications of large language models: An extended comparative analysis. In *International Conference on Human-Computer Interaction*, pages 106–118. Springer, 2024.

144 Chaitanya Sharma. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv:2506.00054*, 2025.

145 Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards anderstanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

146 Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

147 Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. Uncertainty quantification for large language models. In Yuki Arase, David Jurgens, and Fei Xia, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria, July 2025. Association for Computational Linguistics.

148 M Karen Shen and Dongwook Yoon. The dark addiction patterns of current ai chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2025.

149 Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024.

150 Marc Sloan, Hui Yang, and Jun Wang. A term-based methodology for query reformulation understanding. *Information Retrieval Journal*, 18(2):145–165, April 2015.

151 Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

152 Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616. Association for Computational Linguistics, 2025.

153 Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. *arXiv preprint arXiv:2510.11483*, 2025.

154 Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. *arXiv preprint arXiv:2510.11483*, 2025.

155 Zhivar Sourati, Alireza S Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*, 2025.

156 Betsy Sparrow, Jenny Liu, and Daniel M Wegner. Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, 333(6043):776–778, 2011.

157 Matthias Stadler, Maria Bannert, and Michael Sailer. Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386, 2024.

158 Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, 2020.

159 Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, 2024.

160 Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1240–1250, New York, NY, USA, 2025. Association for Computing Machinery.

161 Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14379–14391, 2024.

162 Victor Tangermann. Woman kills herself after talking to openai's ai therapist, 2025.

163 Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843, 2022.

164 Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

165 Robert S Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3):178–194, 1968.

166 Johanne R. Trippas, J. Shane Culpepper, 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, 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 P. De Vries, Ryen W. White, Abby Yuan, Hamed Zamani, Oleg Zendel, Min Zhang, Justin Zobel, Shengyao Zhuang, and Guido Zuccon. Report from the Fourth Strategic Workshop on Information Retrieval in Lorne (SWIRL 2025). *SIGIR Forum*, 59(1), June 2025.

167 Johanne R. Trippas, Luke Gallagher, and Joel Mackenzie. Re-evaluating the command-and-control paradigm in conversational search interactions. In *CIKM*, pages 2260–2270. ACM, 2024.

168 Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2703–2707. ACM, 2024.

169 Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

170 Pertti Vakkari. Searching as learning: A systematization based on literature. *Journal of Information Science*, 42(1):7–18, 2016.

171 V. Venktesh, Mandeep Rathee, and Avishek Anand. Trust but Verify! A Survey on Verification Design for Test-Time Scaling. *arXiv:2508.16665*, September 2025.

172 Ellen Voorhees and Donna Harman, editors. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

173 Ellen M. Voorhees. The evolution of cranfield. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 45–69. Springer, 2019.

174 William Walden, Marc Mason, Orion Weller, Laura Dietz, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, James Mayfield, and Eugene Yang. Auto-argue: Llm-based report generation evaluation. *arXiv preprint arXiv:2509.26184*, 2025.

175 Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*, 2023.

176 Shuai Wang, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Resllm: Large language models are strong resource selectors for federated search. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1360–1364, New York, NY, USA, 2025. Association for Computing Machinery.

177 Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, et al. Memoryllm: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

178 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837, 2022.

179 Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

180 Ryen W White. *Interactions with search systems*. Cambridge University Press, 2016.

181 Joe Wilkins. Leader of albania pelted with trash for appointing ai-powered minister to cabinet, 2025.

182 Lixiang Yan, Viktoria Pammer-Schindler, Caitlin Mills, Andy Nguyen, and Dragan Gašević. Beyond efficiency: Empirical insights on generative ai's impact on cognition, metacognition and epistemic agency in learning, 2025.

183 Wanling Yan, Jialing Li, Can Mi, Wei Wang, Zhengjia Xu, Wenjing Xiong, Longxing Tang, Siyu Wang, Yanzhang Li, and Shuai Wang. Does global positioning system-based navigation dependency make your sense of direction poor? a psychological assessment and eye-tracking study. *Frontiers in Psychology*, 13:983019, 2022.

184 Angela Yang. Lawsuit claims character.ai is responsible for teen's suicide, 2024.

185 Ala Yankouskaya, Magnus Liebherr, and Raian Ali. Can chatgpt be addictive? a call to examine the shift from support to dependence in ai conversational large language models. *Human-Centric Intelligent Systems*, pages 1–13, 2025.

186 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of ICLR*, 2023.

187 Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification, 2024.

188 Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*, 2025.

189 Nadine Yousif. Parents of teenager who took his own life sue openai, 2025.

190 Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 1011–1018. ACM, 2010.

191 İlke Yurtseven, Selami Bagriyanik, and Serkan Ayvaz. A review of spam detection in social media. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 383–388. IEEE, 2021.

192 Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models, 2023.

193 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

194 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

195 Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: anderstanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20, 2023.

196 Tao Zhou and Chunlei Zhang. Examining generative ai user addiction from a cac perspective. *Technology in Society*, 78:102653, 2024.

197 Gabe Zichermann and Christopher Cunningham. *Gamification by design: Implementing game mechanics in web and mobile apps.* " O'Reilly Media, Inc.", 2011.

198 Shoshana Zuboff. The age of surveillance capitalism. In *Social theory re-wired*, pages 203–213. Routledge, 2023.