Report on The Search Futures Workshop at ECIR 2024

Leif Azzopardi
University of Strathclyde
UK
leif.azzopardi@strath.ac.uk

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

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

Bhaskar Mitra
Microsoft
Canada
bmitra@microsoft.com

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

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

Mohammad Aliannejadi, Negar Arabzadeh, Raman Chandrasekar, Maarten de Rijke, Panagiotis Eustratiadis, William Hersh, Jin Huang, Evangelos Kanoulas, Jasmin Kareem, Yongkang Li, Simon Lupart, Kidist Amde Mekonnen, Adam Roegiest, Ian Soboroff, Fabrizio Silvestri, Suzan Verberne, David Vos, Eugene Yang, Yuyue Zhao*

Abstract

The First Search Futures Workshop, in conjunction with the *Fourty-sixth European Conference on Information Retrieval (ECIR)* 2024, looked into the future of search to ask questions such as:

- *How can we harness the power of generative AI to enhance, improve and re-imagine Information Retrieval (IR)?*
- *What are the principles and fundamental rights that the field of Information Retrieval should strive to uphold?*
- *How can we build trustworthy IR systems in light of Large Language Models and their ability to generate content at super human speeds?*
- *What new applications and affordances does generative AI offer and enable, and can we go back to the future, and do what we only dreamed of previously?*

The workshop started with seventeen lightning talks from a diverse set speakers. Instead of conventional paper presentations, the lightning talks provided a rapid and concise overview of ideas, allowing speakers to share critical points or novel concepts quickly. This format was designed to encourage discussion and introduce a wide range of topics within a short period, thereby maximising the exchange of ideas and ensuring that participants could gain

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

insights into various future search areas without the deep dive typically required in longer presentations. This report, co-authored by the workshop's organisers and its participants, summarises the talks and discussions. This report aims to provide the broader IR community with the insights and ideas discussed and debated during the workshop – and to provide a platform for future discussion.

Date: 24 March 2024.

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

1 Introduction

The *First Search Futures Workshop* [Azzopardi et al., 2024], in conjunction with ECIR 2024,[1] aimed to provide a platform for an open discussion about the future of search. To encourage an open discussion in an collaborative fashion, the workshop slides containing the presentations from speakers and breakout groups were shared to create an interactive forum for comment by participants and the community.[2]

The idea for the workshop spawned from conversations during ACM SIGIR 2023 about what the future of search is in light of the generative AI revolution – and wondering how Large Language Models will transform the field. A central question came up, "*Is Information Retrieval, relevant?*". This prompted the need for a forum to discuss and entertain the possible futures of search – and consider the strengths, harms, threats and opportunities that are faced by end-users, designers, researchers, communities and society along with the field itself. While, the workshop was brought about by our deep (and perhaps dire) concerns for the field and its future, where we mused over questions such as: "*How can we trust Gen IR?*", "*What is the point of search, when everything can be generated?*", " *How can we sift fact from fiction?*", "*Will such tools lead to a dystopian nightmare written about in sci-fi novels?*", etc., the workshop was, happily, full of optimism and hope – with many new applications, new affordances and news methods being proposed in light of these developments – as well as the motivating discussions on "*What the field of IR stands for?*" and "*What guiding and underlying principles should we strive to uphold?*". Over 80 participants attended the workshop contributing to the discussions and breakout sessions. By the end of the workshop, despite the many challenges, there was an air of optimism, and sense that there are many new emerging opportunities to research and many open research questions to answer.

2 Vision Statements

During the workshop, presenters shared their views on the future of search – presenting a range of perspectives from seasoned and new researchers and practitioners. Below is summary of their perspectives in their own words. Statements are ordered alphabetically given the first author, while during the workshop (see slide deck) presentations were grouped together into themes regarding applications, theories, and methods – delving into the opportunities along with the potential benefits, harms and threats to users, society and the field of IR.

[1] https://www.ecir2024.org/

[2] https://docs.google.com/presentation/d/1pmA-N1KwoHxAuTWYLkedi-0o8nWQ7rMvyEoFxWk6M64/edit?usp=sharing

Evaluation of Information Access Systems in the Generative Era

Negar Arabzadeh, University of Waterloo

The rapid advancements in information access systems using AI, particularly with the development of Large Language Models (LLMs), have triggered a significant paradigm shift. Traditionally, IR systems focused on retrieving and ranking documents from a static database. However, the current landscape is dominated by Generative Information Retrieval (GenIR) systems, which dynamically generate information from vast, non-deterministic data spaces. This evolution represents not just a technological enhancement but a fundamental transformation in how information is accessed and utilised. Consequently, we need to adapt evaluation strategies that are aligned with these rapid changes and characteristics of new technologies to be able to fairly quantify the utility of information access systems [Arabzadeh and Clarke, 2024a; Alaofi et al., 2024].

GenIR systems introduce complex challenges that fundamentally differ from those encountered in traditional IR. These systems, by their very nature, produce augmented, personalised, and context-aware responses, significantly enriching user interactions. These new attributes would also complicate the evaluation process. Traditional metrics such as precision and recall are inadequate for capturing the effectiveness and alignment of these dynamic outputs with human preferences. There is a critical need to develop new evaluative frameworks that can effectively measure both the qualitative and quantitative aspects of GenIR systems. Recent research has concentrated on developing metrics that assess not only the accuracy but also the creativity and factual correctness of generated content. This effort includes exploring the use of similarity metrics and distributional measures on sparsely labelled data, adapted to evaluate the quality and reliability of the content [Arabzadeh et al., 2024; Arabzadeh and Clarke, 2024b]. Furthermore, there is a growing interest in automating the evaluation process as much as possible to reduce costs and makes it more accessible, and faster [Faggioli et al., 2023; Gilardi et al., 2023; Hou et al., 2023]. As we move toward more automated methods of evaluation for GenIR systems, it is imperative to reconsider the role of human involvement. There is a potential to automate much of the labour-intensive processes traditionally associated with IR evaluation. However, the optimal integration of human judgement remains crucial, particularly in verifying and interpreting the outcomes of these automated systems. In addition, the reliability of these automated labels is still being examined [Ma et al., 2024; Pangakis et al., 2023]. It has been noted that the quality of the labels depends on the calibre of the human annotator. Comparisons between automated labels generated by LLMs and those from human annotators have shown that while they depend on the quality of each, they still demonstrate competitive quality as the automated labels have been recently used in even industrial purposes [Thomas et al., 2024].

The shift towards GenIR systems also raises significant societal implications. The trustworthiness of these systems is not solely dependent on the reliability of indexed items but also on the training data and the algorithms' interpretability. When it comes to evaluation, we need to go beyond just the effectiveness of a system and it is crucial to consider these systems' trustworthiness and how they are aligned with societal values and the ethical dimensions of their deployment [Huo et al., 2023; Maynez et al., 2020; Raunak et al., 2021]. We need to move toward building systems where fairness is quantifiable and correct information is not only available but also identifiable and accessible.

Ubiquitous Finding Engines

Raman Chandrasekar, Institute for Experiential AI, Northeastern University

Today search is typically over large subsets of the web, oriented more at guessing intents across classes of users, and trying to satisfy the vast majority. Personalised search has not really caught on, though there are elements of personalisation in modern search engines. Multi-modal search is not common.

The search engines of the future will focus more on finding than on searching. Searching is more about the process of looking for results appropriate to a given query, while finding is about the outcome and identifying a solution, similar to ideas in [White, 2023]. The metrics of interest for these "*Finding Engines*" will be around task completion and user satisfaction. Workplace productivity increases are not an explicit goal, but highly likely.

Finding will typically be ubiquitous, initiative-taking, federated, and implicit, based on a variety of AI-enabled tools, including but not restricted to LLMs. Finding will be targeted (or focused) in the sense of sub-webs [Chandrasekar et al., 2004], with finding typically focused on specific subject areas, media types, and topical or genre-based subsets of the Web.

Users will usually not have to map their intent to a query and sample results to decide what is relevant. Instead, search will be embedded everywhere, just as electric motors are in our lives today. Computing elements and apps everywhere will use multi-modal context to decide what issues and tasks may be relevant to the user. They will generate queries appropriate to the user's context and federate these queries to appropriate specialist/targeted engines. After filtering away potentially irrelevant responses from these engines, a desired number of override-able action options, resembling recommendations more than results, will be presented to users. User responses and input will be used as implicit and explicit feedback to provide further responses. For instance, the user may suggest variations and what-ifs to elicit further options in this dialog. Multi-modal user-generated queries and options are easy to add to this blend. Visualisation and presentations of responses will be important in this process, determining and demonstrating user willingness to use such systems.

Security and privacy will ideally be balanced to garner user's context in terms of documents and web pages they peruse, their calendar, and interactions on devices to which they have been permitted access. This enhanced context will be used to adaptively generate recommendations/options, trying to avoid being stuck in information bubbles [Slawson et al., 2006]. These options will be ranked not just by relevance, but also on other dimensions. For example, temporal ranking [Chandrasekar et al., 2006] will be used to prioritise options for upcoming meetings, and actions for high-level meetings ranked higher using organisational hierarchy information. Query seasonality and periodicity [Vlachos et al., 2004] could be considered to make it easier to predict interest around specific topics. When so much responsibility is delegated to such systems, trust is key. Attribution and provenance are necessary, and explanations about why each option is offered will help build faith in the system. In addition, users must be trained to understand the limitations of such finding systems, and not to trust them without adequate checks. All this is not easy but is essential for Responsible Finding, akin to similar ideas in Responsible AI [Dignum, 2019].

Several elements of this new search future have been discussed over the years and may be extended with today's tools. It is a successful combination and realisation of these ideas that will be novel. People may still use search as we know it today, say for entertainment or for reference.

Towards Adversarially Robust IR Against Realistic Attacks

Panagiotis Eustratiadis, Yongkang Li, and Evangelos Kanoulas, University of Amsterdam

In 2014, Szegedy et al. [2014] discovered that deep neural networks are vulnerable against imperceptibly perturbed input, termed "adversarial examples", primarily explored within the task of image classification. Here, perturbations are often simple additive noise vectors leading to misclassification. Despite the distinct nature of neural IR from image classification, adversarial IR research largely mimics this paradigm. In IR, documents are subtly altered to impact the ranking outcomes of neural models against specific queries. We identify key flaws in the current adversarial IR paradigm and suggest a future research direction focusing on more realistic and meaningful settings for studying IR adversarial attacks and the robustness of IR models. Our discussion includes attack methods, imperceptibility, and evaluation criteria.

First, we point out that the most feasible method to attack a document corpus is by adding new documents, since existing ones are already encoded and indexed, thus immutable when attacked. For instance, on a social platform, an adversary might seek visibility by creating numerous adversarial profiles to dominate search and recommendation, rather than altering existing profiles. Similarly, adversaries may create fake product pages to outrank established products in e-commerce.

Second, the concept of imperceptibility in adversarial attacks varies significantly across different fields. In computer vision, attacks typically involve adding a noise vector to an image and scaling it down to remain undetectable to the human eye. In contrast, imperceptibility can be achieved creatively in textual content in IR. Techniques include swapping characters for visually similar alternatives (like substituting the Latin letter "*o*" with the Greek "*omicron*"), using very small font sizes, or matching text colour with the background colour. Given these diverse methods, we recommend that retrieval systems be designed to withstand unbounded attacks, ensuring robustness against a wide range of subtle and innovative manipulations.

Finally, we discuss matters of evaluation. In adversarial IR literature, we find the prevalent use of mean rank shift (MRS) or similar metrics to measure the success of an attack inadequate. These metrics judge an attack as successful if adversarially perturbed, yet irrelevant documents consistently outrank their original position across multiple queries. We argue for a more stringent criterion: an attack should only be deemed successful if it significantly impacts normalized Discounted Cumulative Gain (nDCG), meaning that irrelevant documents not only rise in rank but do so above genuinely relevant documents. This tougher standard is not only more challenging but also crucial for robustness, especially considering its implications for retrieval-augmented generation systems that depend heavily on the accuracy and relevance of top-retrieved documents to generate reliable and factual content.

In summary, we believe that while existing work in this field [Chen et al., 2023d; Liu et al., 2023b,c; Wu et al., 2023; Zhong et al., 2023] is both valuable and interesting, future research on adversarial robustness in IR should focus on more challenging and realistic settings. On one hand, we propose making the attacks stronger and unbounded, since imperceptibility is easy to achieve. On the other hand, we propose making the evaluation criteria for successful attacks stricter. Both of these aspects should be considered in the realistic context of polluting the pool of documents with new ones, rather than attempting to corrupt existing documents. We further predict that this

way of looking at robustness will become increasingly important as more and more online content is generated by large language models in a weakly supervised, or even unsupervised manner.

Search still matters in the era of Generative AI

William Hersh, Oregon Health & Science University

Among the hype surrounding generative AI, based on large language models, is the notion that prompting LLMs will replace the need for information retrieval (IR, also known as search). However, based on the state of Gen AI at the time of this writing, there are still important aspects to IR that Gen AI is not quite ready to displace. Let us recall the classic view of IR, formulated by Lancaster [1979] in the last century, declared, "An information retrieval system does not inform (i.e., change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request". In other words, in the classic view, IR systems get people to content items (documents) that may inform them related to their inquiry.

I am not ready to accept the assertion that Gen AI is ready to replace search. From the perspective of the domain in which I work, users have many information needs, from simple to complex [Hersh, 2024]. Users of IR systems, particularly academics, have concerns for authoritativeness – who authored, timeliness – when authored, and contextualisation – larger context of question and supporting evidence. This applies to all of the use cases for biomedical search, including:

- Clinical – patient-care questions
- Research – methods and insights
- Teaching – synthesising knowledge for pedagogy

There is no question that Gen AI has achieved many successes. In the biomedical domain, LLMs have been found to be highly effective in answering clinical questions [Goodman et al., 2023], taking medical board exams [Nori et al., 2023], and solving clinical cases [Tu et al., 2024]. A couple studies have assessed the value of information output by LLMs versus search engines. Hopkins et al. [2023] found that ChatGPT provided more informative information than Google snippets for 4 cancer questions. Van Bulck and Moons [2023] compared the output of ChatGPT vs. Google evaluated by 20 experts in the domains of congenital heart disease, atrial fibrillation, heart failure, and cholesterol. Responses by ChatGPT were deemed trustworthy and valuable, with few experts considering them dangerous.

Nonetheless, as noted above, many users want more than just correct answers from searching. One key output of IR systems is citations so that the searcher can assess the veracity of answers to their questions. A number of studies have found that LLMs fall short when it comes to citations. One analysis of fabrication and errors in bibliographic citations asked ChatGPT to produce short literature reviews on 42 multidisciplinary topics [Walters and Wilder, 2023]. It was found that 55% of GPT-3.5 citations and 18% of GPT-4 citations fabricated and that 43% of real (non-fabricated) GPT-3.5 citations and 24% of real GPT-4 citations included substantive errors.

Another study assessed resource attribution in the biomedical domain comparing several commercial LLMs, one with retrieval-augmented generation (RAG), for their ability to cite relevant references for their claims [Wu et al., 2024]. Clinical questions for prompting were generated

from several well-known Web health information sources. The output was assessed by clinician experts for URL source validity, statement-level support of claims, and response-level support. The best LLM was Microsoft CoPilot (formerly Bing), which includes RAG from the associated search engine. CoPilot had near-perfect URL source validity, 70% statement-level support, and 54% resource-level support. CoPilot had the lowest rate of not citing any references in response to its prompts. Other issues included grounded vs. correct claims and sources behind firewalls.

Clearly Gen AI is going to have a huge role in information-seeking in the future, including in domain-specific areas such as bio-medicine. A big challenge will be methods to insure it makes trustworthy responses, including the citation of references [Coalition for Health AI, 2023]. Researchers will need to develop comprehensive metrics for evaluation [Gienapp et al., 2023] and challenge evaluations like the TREC biomedical tracks will need to assess the use of IR and Gen AI. This will be necessary so that clinicians, patients, researchers, and others can trust the information that comes out of their information-seeking tasks.

IR can help us survive the LLM revolution

Paul Kantor, Rutgers and University of Wisconsin, Madison

In 1986 I proposed (not accepted; no citation) that IR could contribute to Intelligent Systems (as the program was then called). In 1986 Intelligent Systems were generally rule-based and computed logical inferences. As candidate rules proliferated, IR could help to quickly find the ones with the potential to support an inference chain. Before long rule-based approaches had been left the stage; and supporting them seemed unimportant.

Today, in 2024, transformers, and generative statistical models seem poised to once again thwart serious efforts to extract the meanings of texts. Today there is, more than ever, a pivotal role for IR. Here's why: Large Language Models with a billion or trillion parameters ingest "*training texts*". As their Markov models grow deeper, they become ever more able to generate new texts that (because of statistical similarities) have the look and feel of the texts on which they are trained. With a few billion dollars in human tweaking, they respond appropriately to prompts such as "*in the style of Shakespeare*". Amusingly, "*in the style of the New York Times*" is by GPT-3.5 charmingly converted to "*newspaper reading level*" with no nod to the stylistic pretensions of the Grey Lady – currently an odd mixture of confident authority, breathless emotion, and frank click-bait.

The LLM does not understand a word that it is saying. When it creates anew, it is an idiot. Can IR provide guard rails? I believe so. Each LLM educates itself on the accumulated detritus of the WWW as it stands today. As advertisers, journalists, and fraudsters employ LLMs to quickly generate an equal volume of billions of pages with no guardrails whatsoever, they will become exponentially less accurate.

The useful AI(s) of the future should be educated on semi-curated subsets that merit attention. Any (syntactically sound) junk can be used to learn stylistic correlations. The correlations that matter must be learned from texts that mean something, and the something that they mean has to be true, valid, accurate, etc. This has always been the challenge of IR – to find, from among all the indexed and available texts, the ones that will add value to the present state of the seeker's knowledge.

And now for the claim: Soon crucial "*users*" of IR will be LLMs preparing for chatbot careers in specific fields, Medicine, Engineering, Law, Journalism, and of course Politics. For all but Politics their educations will require curated selection of accurate "training documents." Eventually we will debate whether obeisance to "*the already known*" prevents our silicon successors from realising their full artistic and scientific potential. But for the moment our LLMs are in the position of a leader with idiot advisers at both ears. They will lead us to disaster.

When entering the IR field, a bit too long ago, I chuckled at the ease with which some colleagues believed that "*organising existing knowledge*" was the root of all knowledge itself. Anyone who has seen a surprise in a laboratory, through a telescope, or at a geologic site knows how foolish that claim can be. But as we teeter on the precipice of claiming to "*produce new knowledge*" by statistical manipulation of what is already known, sensible and rigorous curation of the "*stuff already known*" has become truly essential.

At a practical level, the IR field could initiate two research tracks: (1) development of curated "*pre-LLM*" pages and resources, to be available for training and (2) research on "*passage level*" provenance certification and tracking.[3] Some combination of blockchain, phrase hashing, and related technologies could be developed and incorporated into the tools used by humans and robots alike, to ensure that a phrase innocently copied from a prior author is automatically tagged and labelled. Not only would this protect university administrators from public embarrassment. It would by its absence, alert readers to the possibility of a LLM confabulation.

Explainability for Engineers: A Path Forward for Search and Recommender Systems in News

Jasmin Kareem, Eindhoven University of Technology, and Maarten de Rijke, University of Amsterdam

Currently, Information Retrieval plays an active role in society. In the news domain, IR systems have the potential to make a large impact on societal debates [Newman et al.; Helberger, 2019]. This impact could be positive or detrimental. Better news search and recommendations could mean a more informed general public, with access to a shared set of facts, potentially leading to better outcomes in general elections. But biased search and recommendations results might lead to the so-called filter bubble [Möller et al., 2018; Stroud, 2011], potentially misinforming individuals, and groups, about the reality behind the news story. Clearly, this outcome depends on how researchers and engineers develop these systems. It is important, therefore, that we understand the capabilities and bear the responsibilities of the search and recommender systems that we produce. And therein lies the problem. In our work with news organisations, we have observed that the transparency needs of the very people who build and maintain the technology to connect people to news are not always met.

Research in explainable AI (XAI) for news often focuses on a single type of stakeholder, the "*user*" [Ter Hoeve et al., 2017]. This is surprising, as previous work has found that in other domains, the majority of deployed transparency mechanisms primarily serve technical stakeholders [Bhatt et al., 2020]. Especially in the news domain, the engineers building the model are key stakeholders with essential transparency needs. Arguably, many XAI methods can be understood

[3]See also: https://www.nytimes.com/2024/01/28/opinion/ai-history-deepfake-watermark.html

by the engineers that build them [Brennen, 2020], but in reality this is not necessarily realised: Kaur et al. [2020] show that data scientists tended to over-trust and misuse XAI tools. This indicates a need for a greater focus on developing methods and tools that target this particular group and that work in practice. Limited progress in improving XAI for news engineers will likely lead to models that display unintended behaviour or that may be evaluated incorrectly. As we move towards a future where large language models are integrated into recommender systems [Di Palma, 2023], general retrieval systems [Tang et al., 2023b], and are used to evaluate the relevance of a retrieved item [Faggioli et al., 2023], the need for having explanation methods that are well suited to be used by the engineers developing and debugging them becomes inescapable.

Given that research in explainable IR is relatively young [Anand et al., 2023], we believe the best path forward towards responsible IR in the news domain, is to focus on improving XAI for engineers to use before deploying an IR system online. This restriction ("before deploying online") is especially important in the news domain: we should avoid using society as a lab. We are optimistic that helping news engineers understand how our search and recommendation models work will lead to models that better serve society's needs. From this perspective, explainability may contribute to solutions to the challenges that IR in the news domain faces.

On the Challenges of Differentiable Search Index in Conversational IR

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas, University of Amsterdam

Two years after the seminal work [Tay et al., 2022] on Differentiable Search Index (DSI), the community has started to address its current limitations [Li et al., 2023b; Ziems et al., 2023; Li et al., 2023a; Chen et al., 2023a; Zhang et al., 2023a; Yang et al., 2023; Zhuang et al., 2023; Wang et al., 2023b; Zhou et al., 2022a; Song et al., 2024]. Recently, thanks to novel documents representations and new training pipelines [Zhou et al., 2022b; Cao et al., 2021; Bevilacqua et al., 2022; Chen et al., 2022, 2023a; Zhang et al., 2023b; Tang et al., 2023a; Zhou et al., 2023; Lee et al., 2023], DSI models can memorize much larger collection of documents [Pradeep et al., 2023; Mehta et al., 2023b]. With this new perspective, a natural question is: *How would DSI impact other fields of research, including conversational IR?*

Conversational IR is the field of research where we retrieve documents based on the user query, considering the conversation context. Leveraging an entire conversation context is challenging as model architectures need to aggregate multi-turn conversations into a single representation. To overcome this challenge, the research community first focused on the rewriting task with encoder-decoder models trained to rewrite the current query into a self-contained one, resolving any anaphora or ellipsis from the context [Aliannejadi et al., 2020; Yu et al., 2020; Vakulenko et al., 2021; Lin et al., 2021; Kumar and Callan, 2020]. However now with the emergence of LLMs that can be used as DSI, and their growing input capacities, such initial solutions need to be revisited, with new training pipelines adapted from the Generative (DSI-QG [Zhuang et al., 2023]) and conversational IR (ConvDR, CoSPLADE [Yu et al., 2021; Hai et al., 2023]) fields.

Another challenge is, while in the vast majority of retrieval-augmented models, retrieval components are trained independently from the answer generation component [Izacard and Grave, 2021; de Jong et al., 2023], several recent approaches now train the retrieval and generation models end-to-end [Yu et al., 2022; Izacard and Grave, 2022]. For example, one solution uses the generation

logit probabilities or attention scores as signal of relevance for the retrieval model [Izacard et al., 2022]. DSI could also leverage such signal, in particular now with the new proposed approach from Zeng et al. [2023] that uses a variation of the margin MSE loss to train DSI. A first benefit of such end-to-end training is that DSI models could be trained without retrieval supervision signal, and only with the weak signal from the answer generation. Then also, it would enable to *train conversational IR models on the entire context at once*, rather than optimising independently for each turn of a conversation. Specifically, this could be done by using a simulated user to follow up on generated answers and only optimise toward full trajectories. This would be promising in term of user intent modelling.

All this considered, it reveals several opportunities for the future of conversational search. In particular on rethinking how to train models to be more aware of a general intent rather than being models that just aggregate or disambiguate queries, but all this comes with modelling and explainabilities challenges.

Beyond the Next Result: Optimizing for Long Term Goals

Kidist Amde Mekonnen, and Maarten de Rijke, University of Amsterdam

Generative Information Retrieval, also referred to as differentiable search index, is a promising approach, employing a unified transformer model for both indexing and retrieval tasks [Tay et al., 2022; Metzler et al., 2021; Tang et al., 2024]. This unified design facilitates end-to-end optimisation towards global objectives and potential integration with reinforcement learning (RL), heralding a future direction focused on long-term retrieval objectives. While promising, GenIR encounters scalability challenges and must adapt to new domains, and understand the nuances of document semantics in docids. However, ongoing research is actively addressing these limitations; see, e.g. [Mehta et al., 2023a; Chen et al., 2023a; Kishore et al., 2023; Zeng et al., 2023; Sun et al., 2023; Chen et al., 2023b; Wang et al., 2023c; Tang et al., 2023a; Zhuang et al., 2023], paving the way for a more robust and versatile GenIR systems. Amidst this evolution, a pertinent question arises: How can GenIR systems be optimised to promote fairness, diversity, and user empowerment, aligning with broader societal objectives in IR? Additionally, What strategies can be employed to design IR systems that reconcile diverse long-term objectives, encompassing user interests, system functionality, and vendor goals?

In search and recommender systems, typical long-term goals focus on extended user engagement, measuring cumulative clicks or dwell time across multiple recommendation sessions [Zou et al., 2019; Deffayet et al., 2023b; Huang et al., 2020; Guo et al., 2016; Gupta et al., 2023; Xu et al., 2020; Zeng et al., 2018; Montazeralghaem et al., 2020]. We propose to concentrate on a different set of goals that better reflect the societal importance and impact of search and recommendation technology:

- Diverse Retrieval: Rather than simply learning a relevance model, we seek to develop models that incorporate different viewpoints and broaden the user's knowledge base. We also seek to provide users with alternative information seeking strategies to support different intentions [Shah and Bender, 2022], and promoting diverse outcomes.

- Fairness-Aware Retrieval: Integrating techniques like de-biasing methods and counterfactual reasoning to mitigate potential biases and ensure fair access to information for users, content providers, and society.
- Multimodal Interactions: Interactions with search systems should be supported through different modalities and different modes, both in terms of user input and in terms of result presentation [Yuan et al., 2024].

Numerous studies emphasise the gap between short-term gains and long-term objectives [Anderson et al., 2020; Hohnhold et al., 2015]. Additionally, traditional search and recommender systems can negatively impact long-term outcomes [Rossi et al., 2021; Deffayet et al., 2022]. Therefore, focusing on the long-term goals of IR is paramount.

Our proposed work aims to leverage RL and extend the optimisation scope within GR beyond short-term accuracy goals. RL has played a pivotal role in addressing sequential decision-making challenges such as query modeling, query expansion, multi-page ranking, and short-term optimization [Odijk et al., 2015; Nogueira and Cho, 2017; Zeng et al., 2018; Wei et al., 2017].

Diverse RL methods like DQN and REINFORCE offer unique strengths. DQN excels in large state-spaces [Roderick et al., 2017] and finds broad application [Chen et al., 2018, 2019b; Huang et al., 2022a,b; Ie et al., 2019b; Liu and Yang, 2019; Liu et al., 2020; Fu, 2022; Zhao et al., 2018, 2020; Zheng et al., 2018]. REINFORCE directly updates policy weights [Chen et al., 2019a; Ma et al., 2020]. Actor-critic methods show promise in handling large action spaces [Dulac-Arnold et al., 2015].

New classes of RL algorithms, such as proximal policy optimisation (PPO) and direct preference optimisation (DPO), offer promise in learning directly from human preferences. PPO's appeal lies in its balanced sample efficiency and simplicity [Schulman et al., 2017; Jain and S, 2019; Ouyang et al., 2022]. In contrast, DPO directly optimises for the policy that best satisfies preferences with a simple classification objective, overcoming the limitations of reward-based methods [Rafailov et al., 2023].

While the current research landscape primarily focuses on applying these algorithms to LLM alignment, exploring their applicability to IR tasks presents a crucial next step. This exploration would require tailoring these algorithms to address specific IR challenges and evaluating their effectiveness in real-world settings. Such explorations could pave the way for a new era of user-centric IR, where results not only meet relevance criteria but also align with the higher-level goals formulated above.

A recent study employing RL within the context of GR addressed the challenge of misalignment between token-level optimisation and document-level relevance, as well as the tendency to over-emphasise top-ranked results [Zhou et al., 2023]. This work provides a valuable foundation for further exploration. Two key areas for future research emerge from this study. First, we can investigate the applicability of DPO to short-term scenarios. This exploration would expand the range of techniques available for facilitating learning that aligns with long-term objectives. Second, we can extend this approach beyond accuracy-focused goals. Our future focus should shift towards holistic, long-term objectives exceeding mere accuracy metrics. This necessitates exploring alternative evaluation paradigms and designing IR systems that reconcile diverse priorities, including user interests, system functionality, and vendor goals. By addressing these questions,

we envision building a robust and adaptable IR system that strikes a balance between short-term efficiency, long-term relevance, and user satisfaction.

Re-centering Search on Societal Needs

Bhaskar Mitra, Microsoft Research

The field of Information Retrieval is currently undergoing a transformative shift. Large language models (LLMs) are rapidly changing how we access information by introducing new information access modalities (*e.g.*, Bing Chat[4] and Google Bard[5]) and by embedding themselves in user's work processes where they interact with the IR system on the user's behalf (*e.g.*, Microsoft Copilot for M365 [Mehdi, 2024; Warren, 2024]), under retrieval-augmentation [Lewis et al., 2020; Zamani et al., 2022]. LLMs have also demonstrated incredible improvements in automatic relevance estimation, a core IR task, demonstrating the ability to estimate the searcher's preferences better than several populations of human relevance assessors [Thomas et al., 2024]. While these progresses are reasons for genuine excitement in the field, it is also noteworthy that many of these developments are coming on the backs of fundamental research being conducted in IR-adjacent communities, such as machine learning (ML) and natural language processing (NLP), leading to some seeing IR as just another NLP task. It is in this moment of both uncertainty and opportunity that the field must articulate an aspiring vision for IR research, or risk being reduced to just an application of ML and NLP. We believe that as part of articulating this aspiring vision, IR should re-center itself on the needs of society.

Information access plays a critical role in society in several ways including by shaping political discourse, supporting public health education, and aiding knowledge production. IR research must therefore both understand and contend with the societal implications from the technology it produces. This is not a new perspective and has been echoed by Belkin and Robertson [1976] in the IR community nearly half a century ago. More recently, there has been several calls [Culpepper et al., 2018; Olteanu et al., 2021] for increased investment in research directions of social import, such as fairness, accountability, confidentiality, transparency, and safety in IR, followed by a large body of recent work [Ekstrand et al., 2021; Zehlike et al., 2022a,b; Pitoura et al., 2022; Dinnissen and Bauer, 2022; Aalam et al., 2022; Wang et al., 2023a; Li et al., 2023c; Deldjoo et al., 2023; Zhang et al., 2020; Anand et al., 2022] in these areas. This burgeoning focus on fairness and ethics in IR has brought much-needed attention to the IR community's responsibility to broader society. However, this approach has also been largely reactionary trying to mitigate potential social harms from emerging technologies by developing new fairness and transparency interventions, which operates within a severely constrained frame that has left many underlying assumptions about the sociopolitical and economic incentives that guide IR research largely unchallenged.

Instead, it is our perspective that the field should aim to proactively set the research agenda for the kinds of systems we *should* build motivated by broader societal concerns and aspirations and by challenging the economic and political power structures within which we conduct our research. These considerations should be central to all research in our field and we should dismantle the artificial separation between the work on fairness and ethics in IR and the rest of IR research.

[4] https://chat.bing.com/

[5] https://bard.google.com/

Besides the users, the content producers, and the system owners, IR research should explicitly recognise society as a stakeholder in the context of system design and deployment. We believe that IR research should explicitly reflect how these systems should contribute to knowledge production, pubic education, and social movements, and that broader framing of societal concerns should be the most fundamental stakeholder need that should inform and shape IR research. Towards this goal, IR research must also break out of its silo and engage with social scientists, legal scholars, critical theorists, activists, and artists, in a recognition of a collective struggle towards a better future for all.

Note: The above statement is based on a broader perspective titled "*Search and Society: Re-imagining Information Access for Radical Futures*" [Mitra, 2024].

Going Back to The Future

Adam Roegiest, Zuva

With the arrival of ChatGPT [OpenAI, 2022], many researchers have begun to ask themselves "*where should we be going?*". In contrast, one of the interesting questions that does not often get asked is "*where did we want to go in the past but couldn't?*". Understanding the IR community's shared history can help provide old insights to current problems, whether this is due to old results being more relevant or bottlenecks that impeded process.

As the workshop asks "*is there a place for the guiding principles of IR, in this brave new world?*", a reasonable counter would be "*does any one even know what the guiding principles of IR are?*". There seems to be no canonical written history of these principles of IR nor does there appear to the author to be a shared understanding of what they were, are, or should be. As long-standing members of the IR community slowly depart, we begin to lose the ability to understand the historical principles that shaped the growth and development of IR and the community is tasked with inferring, potentially incorrectly, what they were from published literature.

For established members of the community, there may be some intuitive knowledge of current principles. But with the influx of new members to Information Retrieval, spurred by generative models, from adjacent communities (e.g., Recommender Systems, NLP, AI/ML) and a continual stream of new students, how do we expect these individuals to come up to speed in as smooth as process as possible (e.g., without having to infer from reviewer comments what the IR community values). By formalising these principles, we can help guide new community members on how best to contribute and potentially provide an understanding of why the IR community values the things that it does.

At the same time, this is not a proposal to have every IR researcher know the precise details of the past but that we should have mechanisms to more easily explore that past. After all, we are Information Retrieval experts. One such avenue, could be through reviving the "*Readings in IR*" [Jones and Willett, 1997] or through community-historians and workshops or maybe an IR chatbot. In building a shared understand of IR's history and its ever elusive guiding principles, we may continue to find identity in exploring the core IR problem of "*given a user and their information need, how do we satisfy it?*".

Information Retrieval in the Age of RAG Systems

Fabrizio Silvestri, Sapienza University of Rome

Information Retrieval is on the brink of a transformative era influenced by the integration of Retrieval-Augmented Generation (RAG) systems. RAG systems have emerged as a significant technological ad-

vancement, primarily due to their ability to enable large language models to generate updated and accurate responses. A notable potential benefit of these systems is their capacity to mitigate the issue of "hallucinations," where the model generates incorrect or misleading information.

One of the central questions facing the field of IR is whether current systems are adequately equipped to support the functionality of RAG systems. This question extends to the suitability of dense versus sparse IR techniques and whether a multi-collection IR or a mixture of IR experts' approaches could enhance the efficacy of these systems.

Moreover, the relationship between RAG systems and LLMs invites further scrutiny. It is crucial to consider whether LLMs, as they are currently developed, are optimal for integration with IR systems or if they need to rethink their design specifically for RAG applications. Additionally, the role of "noise" or unstructured data in improving the performance of RAG systems suggests a potential reevaluation of data usage and system training in IR.

Inspired by the study done by Liu et al. [2024], where it is shown that the position of the retrieved relevant results impacts the overall quality of the result generated by the LLM of a RAG system, we propose in [Cuconasu et al., 2024] to investigate the role of "noise" or unstructured, irrelevant data in improving the performance of Retrieval-Augmented Generation (RAG) systems. This examination underscores a counter-intuitive finding: including seemingly irrelevant documents within the information retrieval (IR) phase can significantly enhance the system's accuracy by more than 30% in some scenarios. Contrary to initial assumptions that such documents would degrade the quality of outputs, noise introduces a beneficial complexity that aids in reducing errors typically associated with LLMs, such as hallucinations or generating factually incorrect information.

This phenomenon suggests that the standard approaches to optimising IR systems for accuracy and relevance might need reevaluation. Traditional metrics prioritise retrieving relevant documents to improve precision. However, the study illustrates that introducing a controlled amount of irrelevant information can paradoxically make the RAG systems more robust. The noise enriches the model's context, providing a broader spectrum of information that prevents the model from overly focusing on potentially misleading cues in highly relevant documents.

Further research will need to explore the optimal level and types of noise that can be introduced without compromising the system's functional integrity. Moreover, these findings challenge the current understanding of data quality in machine learning, potentially leading to innovative methodologies in both the development and evaluation of RAG systems. Looking ahead, the interaction between IR and RAG systems is important to consider. One such consideration is the efficacy of prompting in facilitating the interaction between IR systems and LLMs. The possibility of moving towards an end-to-end training model for RAG systems could represent a significant shift in how these technologies are developed and implemented.

In conclusion, the future role of IR in retrieval-augmented systems is poised for significant evolution. As these systems become increasingly sophisticated, the need for advanced IR solutions to effectively support and enhance RAG systems will be critical. The ongoing development of these technologies will likely focus on improving the compatibility of IR systems with LLMs, optimising system architectures for enhanced performance, and refining the methodologies used to integrate and train these complex systems.

Trustworthy Information Systems

Ian Soboroff, National Institute of Standards and Technology

At the beginning of 2023, NIST published an AI Risk Management Framework [NIST, 2023]. This document is intended to help companies and organisations identify risks from the use of AI. If the risks

can be identified and mitigated, that is said to improve the *trustworthiness* of the system in question. However, the document does not define trust or address its dynamic nature.

Wikipedia defines trust as "*the belief that another person will do what is expected ... in ways that benefit the trustor. In addition, the trustor does not have control over the actions of the trustee.*"[6] Trust is an active area of research in many social sciences, and rather than plumb those rich depths I will stick with the above. I further take "*another person*" to be anything that can act on behalf of a person. Shah and Bender identify several contributors to trust and trustworthiness in information systems and give examples of situations which might affect trust [Shah and Bender, 2024].

Trust is a belief with magnitude: we can have more or less trust, and trust can be gained or lost due to the actions of the trustee. Even though its definition is elusive, we can perceive whether we trust another and changes in that degree of trust. We say that someone or something "earns our trust". Likewise, trust is testable. A friend earns my trust and is then deemed trustworthy; a betrayal of that trust is a failed test which reduces trust, while other actions might increase trust.

Information systems can earn trust by being reliable, supporting the user's task, being transparent in their operation, and exhibiting repeatable (and hence predictable) behaviour. Surprises or contradictions of held beliefs are a test of trust: they reduce trust unless the user can convince themselves or be convinced to change their prior belief. Convincing and validated revisions of prior belief increase trust. So, in order to pass the test, the system must support this process of convincing through a "*dialog*" with the user.

The traditional "*ten blue links*" results view from a search provides a number of affordances that support this dialog. Each hit provides a title and a snippet, and from experience we trust those extractions to have come from that document (but see [Shah and Bender, 2024] for examples of noted failures here). The hits are in a ranked order, and from experience we trust that the system is at least trying to order results such that the most helpful or relevant are at the top. Each hit shows a URL, the source of the information, and we trust that those URLs are correct again from experience. We are able to make judgements of the quality of the hit from all these cues. We can revise the query and with experience this affects the results in more-or-less predictable ways. In summary, experience provides a level of trust, and the results display provides several ways in which that trust can be tested. One important trust gap comes from a lack of evidence for recall; even when the searcher needs only a single answer they might not be aware of a better one.

A generated response to a search query only supports the trust process through the text itself. While current technology can be made to mimic a multi-turn conversation, the model has no understanding of the context of the information and so is not able to produce an explanation of its behaviour that a user could make use of. A retrieval-augmented generation process can help by providing citations, which the user can explore to verify the text. A citation on a generated passage is different from a snippet which is assumed to be extracted directly from the source document. Then, based on the citations, a user could form a prompt that takes them deeper into the model's "*grasp*" of the information. This is a lot of effort for a user to go to, but these tools are yet young and most people are not experienced with them.

Of course, since the interface described is purely generative, it's turtles all the way down: *if you don't trust one output, why should the next be even as trustworthy?* This is another difference from the traditional search engine interface: beyond the ten blue links are the web pages themselves.

Information Retrieval as the Trustworthy Alternative

Suzan Verberne, Leiden University

[6] https://en.wikipedia.org/wiki/Trust_(social_science), visited May 9, 2024.

Large Language Models are capable of generating fluent language. This offers potential for interactive and tailored interaction with a user. On the other hand, text generation has the risk of hallucination [Ji et al., 2023] and aggravation of bias [Navigli et al., 2023].

Therefore, two important pillars in the future development of LLMs are humans and sources. Humans are needed to verify the accuracy of information, and for this purpose, the source of the information is necessary. Sources enable us to assess the quality and value of the information. These two aspects, the human and the source, are precisely why search engines are so valuable: when we ask a search engine a question, it will show us a list of sources found on the web. The information is not "*hallucinated*" but retrieved from the index. It is up to the human to assess the value of the information. A search engine is inherently "*human in the loop*": the human determines the search query and selects the relevant information.

By showing sources, search engines provide a form of transparency. Snippets indicate to the user why the document was considered relevant to the search query, by bold-facing query terms. Search engines also enable us to explicitly incorporate representativeness and diversity into the results [Santos et al., 2015; Abolghasemi et al., 2024].

As pointed out in the previous sections, retrieval can help LLMs through Retrieval-Augmented Generation (RAG) [Lewis et al., 2020]: first retrieving relevant documents from an index, then giving them to the LLM to formulate the answer based on these sources. This is useful in many situations: to have access to recent information that was not in the pre-training data of the LLM, or to have access to organisation-specific or personal information that is not part of the LLM pre-training data.

On the other hand, LLMs can also be of service to retrieval models: to generate data with which we can train rankers. If there are documents available for a task but no search queries, we can use an LLM to generate potentially relevant questions for each document [Jeronymo et al., 2023]. Conversely, if we have search queries but no relevant documents, we let a LLM generate the relevant text passages [Askari et al., 2023b,c]. We then use the pairs of search queries and relevant documents to train a ranking model.

The risk of training retrieval models with hallucinated information is a potential concern of this approach. In my view, however, the risk of hallucination is a reason to opt for this indirect use of generative language models: as data generators instead of direct use to answer questions. After all, a search engine will still only retrieve information available in its index, and not hallucinate. Moreover, when we generate training data with an LLM, we can artificially introduce a balance in the training data that is not present in natural, human-generated data, thereby reducing bias in the model [Abolghasemi et al., 2023].

There are some tasks for which we want to steer away from the fluent interaction that LLMs have to offer, i.e. in domain-specific question answering (QA) settings where trustworthiness is. Key examples are legal and medical QA where a layperson asks a question to an expert [Askari et al., 2023a]. In these settings we need retrieval to ensure the reliability and trustworthiness of the returned information.

The Role of LLMs in Democratic News Rec Sys

David Vos, Jin Huang and Maarten de Rijke, University of Amsterdam

News recommender systems are an opportunity to advance values and goals in a democratic society [Helberger, 2019]. However, only optimising these recommender systems for user engagement potentially has negative implications for democracy, because of filter bubbles and polarisation, among other things [Pariser, 2011]. To align recommender systems with democratic goals, a notion of diversity is required that accounts for norms and values [Helberger, 2019]. While the normative understanding of diversity has been extensively studied in social science literature [Loecherbach et al., 2020], only limited attention has been devoted to normative diversity in recommender systems.

Helberger [2019] provides a conceptualisation of normative diversity within the context of democratic goals and explores how news recommenders can contribute to these goals. Building upon this foundation, quantitative metrics have been defined for evaluating recommendations in alignment with normative goals [Vrijenhoek et al., 2021, 2022]. These approaches heavily rely on the expertise of individuals and lack the agility to swiftly adapt to dynamic changes in the world and user preferences.

We propose to use the capabilities of large language models to integrate notions of normative diversity into democratic news recommender systems. We hypothesise that, compared to manually defining metrics of normative diversity, LLMs can do this more dynamically, based on subtle textual differences and changes in content.

When exploring this application of LLMs, we propose an emphasis on research that helps an LLM capture the norms and values of a news outlet. Furthermore, research should focus on giving journalists as much control over the recommendation strategy as possible. We propose two main lines of research that combine these two factors.

The first direction of research should be the exploration of fine-tuning strategies. By fine-tuning an LLM based on news articles published by journalists, it could potentially better capture the details of a journalist's writing. By designing expressive fine-tuning strategies, the subtle details of normative diversity can be captured increasingly well.

Secondly, to give journalists even more explicit control, we propose to adapt LLMs through reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022]. By using RLHF, journalists can influence the recommendation policy by providing explicit feedback on recommendation sets. Research should focus on designing RLHF pipelines that are highly expressive but intuitive for journalists.

When working on LLMs for normative diversity, it should not be forgotten that LLMs are black-box algorithms that have been shown to manifest social biases [Liang et al., 2021]. Considering these issues is especially important when using LLMs to assess nuanced topics such as normative diversity.

Recent work has attempted to translate the definitions of norms into practical diversity metrics [Vrijenhoek et al., 2021, 2022]. However, the news domain is highly dynamic, and norms are hard to capture in static distributions. We believe that research should focus on evaluation methods through user studies with journalists, or by learning from user interactions in production environments.

In this work, we have proposed several research directions into using LLMs for integrating notions of normative diversity in news recommender systems. We hope future work will continue research into developing and evaluating these methods.

IR for Complex Information Needs

Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University

Typical ad-hoc search engines nowadays provide the user with a ranked list of documents or links based on short keyword queries entered by the user. Under this setup, the community developed lines of approaches, such as pseudo-relevance feedback and query intent classification, to improve the effectiveness of the search results without additional explicit user input other than queries. However, with the recent introduction of generative models with a chat interface, people argue that generative systems will replace search engines because of their ability to handle complex information needs. I argue that these generative model-based systems are still driven by the same information needs that drive the development of search engines.

Users developed their search workflows to work with existing ad-hoc search engine systems by breaking complex information needs into steps of simpler queries that the system could process. While effective, reading through lists of web pages is not always the most effective and efficient result presentation for human consumption [Hearst, 2011]. With generative models and retrieval augmented generation, we have

the potential to process complex inquiries directly and provide rich and informative responses. In both cases, users are driven by the same information need but allocate their effort in different ways.

Assuming one-off ad-hoc queries is a convenient (and useful for a long time) assumption for developing retrieval systems. Works in session search [Kanoulas et al., 2010; Jiang et al., 2014; Liu and Belkin, 2010] began to recognise the continuation of search queries that users leverage to gather relevant documents for more complex information needs; high recall retrieval [Baron et al., 2016; Harty, 2017; Wallace et al., 2010] also spawns a line of work on assessing the effectiveness of workflows instead of one single search [Yang et al., 2021]. Generative models provide an alternative approach for users to express their needs directly instead of manually decomposing them and combining the results. Although applications with generative models are still in their early stages, the fact that users actively engage with systems and provide a long description of the task in exchange for rich, complex system responses demonstrates the potential of designing systems expecting more interactions between the user and the system.

Furthermore, the presentation of the results should play an important role moving forward. The "ten blue links" presentation drives the development of ranked retrieval in many aspects, including evaluation measures, rank fairness, re-ranking approaches, etc. It is the presentation that propagates back to the algorithmic developments. Exploring different result presentations will create even more opportunities in all different aspects of information retrieval.

TREC Complex Answer Retrieval track that ran from 2017 to 2019 [Dietz et al., 2017, 2018; Dietz and Foley, 2019] is an earlier attempt to serve complex information needs and is perhaps ahead of its time. It aimed to develop systems that would take in simple user queries and provide a rich, long response covering a range of aspects. Since generative models were far less capable, such visions were difficult to materialise in real systems.

With the current technology, it is possible to generate informative responses tailored to a specific user. In TREC 2024, three tracks, including NeuCLIR, Retrieval Augmented Generation, and Biomedical Generative Retrieval tracks, were proposed to evaluate system-generated responses. Especially in NeuCLIR,[7] the organisers[8] propose a report generation task [Mayfield et al., 2024; Barham et al., 2023] that takes a paragraph of a user report request, searches a fixed document collection, and synthesises an informative report with citations based on the retrieval results. This task is an example of exploring alternative result presentation and aims to drive new retrieval modelling and algorithmic research.

Information retrieval, rather than in its dusk, is in a very exciting era. By redefining what a retrieval system can and should do, we can further satisfy the end user with more results that are more informative to the specific user and easier to consume. Such transitions will foster more research and better systems in the long run.

Can LLMs Serve as User Simulators for Rec Sys?

Yuyue Zhao, Jin Huang, and Maarten de Rijke, University of Amsterdam

During the development of recommender systems (RSs) researchers and practitioners often rely heavily on logged user interactions for training and evaluation [Oosterhuis and de Rijke, 2021]. However, the reactions of real users are often complex, and logged interactions are limited, e.g., by the logging policy, resulting in performance degradation when interacting with actual users [Chen et al., 2023c; Gilotte et al., 2018]. An alternative is to use simulation-based experiments, which simulate user feedback on recommended items. Simulators can easily scale up to cover a range of scenarios, allowing one to test the adaptability of RSs [Bernardi et al., 2021]. Various user simulators have been proposed and used to

[7] https://neuclir.github.io/2024

[8] The author of this statement is one of the organisers.

develop RSs; see, e.g. [Deffayet et al., 2023a; Huang et al., 2020; Chen et al., 2023c; Ie et al., 2019a; Shi et al., 2019]. Ie et al. [2019a] provide a platform for authoring environments to simulate user feedback on sequential recommendations. Huang et al. [2020] and Chen et al. [2023c] create simulators to support the optimisation of reinforcement learning-based methods for long-term engagement.

While simulators have led to significant performance improvements in some scenarios [Huang et al., 2020; Chen et al., 2023c], we recognise two important limitations: (1) They have a limited ability to integrate diverse data sources, including various types of user behaviour (e.g., ratings, reviews, images, speech) and heterogeneous user and item contexts. (2) They ignore the sequence of cognitive and emotional processes that users experience when expressing their preferences, and do not perform self-refinement as a human would in decision-making [Simon, 1962]. We propose to build a user simulator that is more adaptable to diverse recommendation tasks and can model more human-like user responses.

Large language models have gained attention in part due to their success at NLP tasks [Li et al., 2024; Liu et al., 2023a; Dubois et al., 2024] as a result of pre-training on extensive datasets and access to open-world knowledge [Achiam et al., 2023]. LLMs have been used to simulate human behaviour, attitudes, and emotions [Gao et al., 2023], and can refine their decisions based on self-provided feedback [Madaan et al., 2023]. Inspired by these findings, we propose to use LLMs as user simulators for RSs. Our approach configures LLMs as agents to simulate user types, incorporating a reflection strategy to infer users' latent preferences and refine the simulation process for more accurate user simulation.

Using LLMs as user simulators for evaluating RSs presents three main challenges. The first involves enhancing LLMs' recommendation capabilities to more closely mimic actual user responses. The second is to ensure that simulated behaviours reflect users' historical interactions and current contexts, considering that user preferences may vary with circumstances. The third challenge is developing reliable methods to evaluate the effectiveness of these simulators, especially given the "black box" nature of LLMs.

Currently, our focus lies on fine-tuning methodologies, prompt engineering, and reflection strategies to improve the accuracy of simulated users, as well as constructing benchmarks to evaluate user simulators performance. Moving forward, we aim to minimise hallucinations in the user simulation process, and to address ethical concerns such as fairness and bias. We believe that LLM-based user simulation will offer a safer, more cost-effective alternative to conventional online A/B testing in the future.

3 Breakout Group Summary

During the afternoon, participants split out into several breakout groups to discuss areas of interest involving: trust, evaluation, simulation, new applications and societal impacts. Below is summary of some of the discussion points from various breakouts.

3.1 Trust

One important opportunity for defining the future of IR lies in the notions of trust, trustworthiness, and truth. Until the advent of LLMs, most of the material available on the Web was either written by human agents, or was recognisably botogenic, due to the simplicity of artificial generation. Accelerating improvement of LLMs make it increasingly difficulty, if not impossible, to distinguish human-generated from machine-generated texts by examination of the texts themselves. Similarly, the idea that one might authenticate a source by conversation with the "*author*" is impractical because humans do not have the bandwidth to respond to every inquirer, for which LLMs are increasingly able to provide plausible, if anodyne, responses to reasonable inquiries.

Trust is a relation between two agents. In this setting it is between a human user, and an "*information item*" — which may be a text, a database, or tutorbot, etc. In our conversation we recognised that one cannot compel a given user to trust only that which is true. For example, a believer in flat earth will not accept the kinds of evidence that most of us accept. The relation between the system and the user also recognises at least two ways in which users approach systems. One, let us call it "*hedonic*" is for the pleasure to be had from using the system. Examples range from cat pictures to flat earth theories. The second, let us call it "*factual*" is for the knowledge to be gained from using the system. Examples range from booking a train ticket to understanding Bayesian networks. Generally speaking, service to the factual user is better if the information delivered is promptly accessible, readily comprehensible, and true to the world "*as it really is*". Service to the hedonic user, by contrast, is better if it provides an engagement of the desired duration, and induces the desired response, which may be pleasure, anger, or relaxation. Insofar as the system can determine which type of user is at hand, the appropriate metrics can be applied in the laboratory and the real world. If the match is good, trust will grow and have been well earned.

There is a troubling middle ground, which we do not address, in which the user is encouraged to perform actions that harm to themselves or others. Examples include some portrayals of eating disorders, alternative beliefs affecting public health – AIDS, measles, and so forth. This is a risk to both kinds of users: the hedonic user has a low expectation of risk (because of a lack of engagement), while the factual user may be overly inclined to trust the information they find (because of cognitive biases).

Roughly speaking, the hedonic and the factual users are oppositely served by the products of LLMs. To the extent that LLMs can mimic the desired features of materials sought by hedonic users, they free human beings from the burden of creating it, and can let them move on to other activities. While this will lead to displacements in such domains as pornography, and perhaps eventually humor, we choose not to regard it as a problem.

In sharp contrast, the factual user may be harmed by LLM products, precisely because there is nothing in the design of generators that constrains what they produce to be true or correct. And, to the extent that their products are added to the training resources, there is an alarming potential for a disastrous feedback loop in which truthful information is buried beneath an ocean of confabulation. Thus it seems essential to augment existing methods of IR to allow for these two classes of users, and for the fact that generated materials may "*look indistinguishable from*" the materials used to train the generators.

It seems essential that the IR system know more about the materials that it indexes, so that it can answer the relevant questions from either class of user. A shopper, for instance, may want to know how many others have purchased a given item. A researcher may want to know how much a paper has been cited, whether it has been withdrawn, where its authors did the reported work, etc. A surfer may want to know whether an item has been endorsed by a favourite singer, or athlete, or politician. Etc. There is clearly too much that can be known, and it can not be front-loaded, where it becomes an obstacle to the user.

As with information itself, these metadata may be treated as either a filtering or an *ad hoc* problem. Treating it as *ad hoc*, the user should have an easy way to query the metadata, for those dimensions that matter to that particular user. A typical use case, with traditional retrieval, is that the user finds a snippet interesting, but before proceeding wants to check on its date of creation. This might be a "*two-click*" process – one to get to the metadata selector, and a second to ask for the data such as the data (or hour) of creation of the object. We recognise that this will require solutions to a number of open problems in data analysis. As an example, some search engines add a "*time stamp*" such as "*2 hours ago*" to a news snippet, and one cannot know whether the snippet itself is only 2 hours old, or that someone has added a comma somewhere in the text. Overall, these are the kinds of problems that

IR has been dealing with since its inception, and there is every reason to believe that IR researchers will contribute crucial insights, techniques and principles to the eventual resolutions.

Another perspective on trust is that trust is earned and testable. A friend earns my trust and is then deemed trustworthy; a betrayal of that trust is a failed test which reduces trust, while other actions might increase trust. Information systems can earn trust by being reliable, supporting the user's task, exhibiting repeatable (and hence predictable) behaviour. Surprises or contradictions of held beliefs are a test of trust: they reduce trust unless the user can convince themselves or be convinced to change their prior belief. Convincing and validated revisions of prior belief increase trust. So, in order to pass the test, the system must support this process of convincing through a "*dialog*" with the user.

The traditional "*ten blue links*" results view from a search provides a number of affordances that support this dialog. Each hit provides a title and a snippet, and from experience we trust those extractions to have come from that document. The hits are in a ranked order, and from experience we trust that the system is at least trying to order results such that the most helpful or relevant are at the top. Each hit shows a URL, the source of the information, and we trust that those URLs are correct again from experience. We are able to make judgements of the quality of the hit from all these cues. We can revise the query and with experience this affects the results in more-or-less predictable ways. In summary, experience provides a level of trust, and the results display provides several ways in which that trust can be tested. One important trust gap comes from a lack of evidence for recall; even when the searcher needs only a single answer they might not be aware of a better one.

A generated response to a search query only supports the trust process through the text itself. While current technology can be made to mimic a multi-turn conversation, the model has no understanding of the context of the information and so is not able to produce an explanation of its behaviour that a user could make use of. A retrieval-augmented generation process can help by providing citations, which the user can explore to verify the text. A citation on a generated passage is different from a snippet which is assumed to be extracted directly from the source document. Then, based on the citations, a user could form a prompt that takes them deeper into the model's "*grasp*" of the information. This is a lot of effort for a user to go to, but these tools are yet young and most people are not experienced with them.

Of course, since the interface described is purely generative, it's turtles all the way down: if you don't trust one output, why should the next be even as trustworthy? This is another difference from the traditional search engine interface: beyond the ten blue links are the web pages themselves.

3.2 Evaluation and Simulation of Interactive IR

Large Language Modelling and Generative AI/IR methods provide a whole raft of new possibilities for the evaluation and simulation of Interactive IR systems.

In terms of evaluation, it was felt that GenIR breaks a number of assumptions invoked by the Cranfield Paradigm – and that a more emphasis needs to be placed on evaluating the responses given an information need, rather than the relevance of documents. Citing the report by Gienapp et al. [2023], discussions focused around how and what we are evaluating is changing. In particular, how the coherence, coverage, consistency, correctness and clarity of responses and statements made by Generative IR approaches needs to be considered within evaluations taken against the costs (e.g., computational, temporal, fiscal and environmental). Moreover, evaluation paradigms need to evolve from the evaluation of static rankings to consider multi-turn interactions – where we need to move beyond evaluating the ranked lists to focusing in on the statements, nuggets and references contained in responses – and do so in the context of interaction and conversation. This significantly increases the complexity of evaluations that is further compounded by considering the broader array of tasks that could be addressed by search

engine that utilise retrieval augmented generation. This signals a shift towards evaluating task completion and finding engines rather than traditional search engines.

In terms of simulation, it was felt that with LLMs we could simulate almost anything – and evaluate more deeply interactive and conversational systems – across a variety of simulated personas. For example, among other things, we could and can generate:

- labels, annotations, judgments, etc.,
- answers, responses, documents, collections, etc.,
- explanations, descriptions, reasons, recommendations, etc.,
- interactions, clicks, likes, feedback, etc.,
- user intents, queries, needs, tasks, etc., and,
- user states, personas, objectives, goals, etc.

Already there has been much interest in using LLMs to generate judgments in order to evaluate system rankings to great success (e.g., [Faggioli et al., 2023; Gienapp et al., 2023; Awasthi et al., 2023; Sakai, 2023; Farzi and Dietz, 2024; Wu et al., 2024]). Clearly, LLMs offer a way to create labels fast and cheaply and to do so in a various settings. So, what if we took this idea further, to generate and produce other the types of data. Could we build "*simulated users*" that realistically mimic how people of different backgrounds and demographics with different needs and intents would interact with different (conversational) IR systems? In a previous workshop on the simulation of interaction held at ACM SIGIR over ten years ago, such things were only dreamt of [Azzopardi et al., 2011]. Now, we have the potential capabilities to perform much richer and nuanced simulations of interaction. From these we can create repeatable and reusable simulated users to help evaluate systems and to provide useful insights that inform the design and development of future IR systems. These simulated users might even evolve to become agents that search and work on our behalves.

Discussions on these ideas led to further discourse over a variety of topics, for instance: rather than developing static test collections, per say, we could shift to developing test simulated users, that could repeatedly and reliably be used to evaluate current and future systems. This could help overcome the central challenge of evaluating any interactive system – the infinite state space due to interaction [Belkin, 2008]. If we move to simulated users and agents, then how do we seed and ground them to be representative of actual users? If these agents are then used outwith evaluation, but as a tool for real users, then how can/should we interpret or filter the signals and data produced by these agents? And, what if this data is used to train subsequent agents, what issues might arise (e.g., lack of diverse interaction, accumulation of biases, etc.)? And when creating all these new simulated users, how would we validate and evaluate these users? And what if they could also generate content along with their interaction data – how could this change the information space. Could such agents be used to attack, misinform, spam, etc.? And as the simulated agents become more and more realistic, how would we be able to tell the difference between carbon and silicon LLMs?

During the breakout session, participants also found that there was a quorum of researchers interested in simulating users for various purposes – and that they had been using SimIIR, a framework for the Simulation of Interactive IR [Maxwell and Azzopardi, 2016; Zerhoudi et al., 2022]. This chance meeting of participants led to further discussions during ECIR's Collab-a-thon and the creation of a new GitHub Organization[9] for building simulated users and agents. A channel `#simiir`[10] on ACM SIGIR's Slack was also created to help facilitate the ongoing collaboration where all are welcome to join and participate.

[9] https://github.com/simint-ai

[10] https://acmsigir.slack.com/archives/C06RC0A31EX

3.3 New Applications and the Applicability of GenAI

A large component to understanding where to go is understanding where we are and where we have been. A large part of the discussion in this breakout group was in situating current Information Retrieval systems and practices using the two questions: "*why is current search good enough?*" and "*how is the current search good enough?*" By discussing historical insights into systems as well as contemporaneous issues in e-commerce systems (e.g., website specific jargon impeding search success), the group sought to understand what areas need new developments and what ones might be good enough as is to not warrant further attention.

Helping user's to query effectively was an actively discussed topic. As search technologies have largely shifted how user's query for information (i.e., short keyword-based queries). This began with a discussion of how this had been explored in conversational search research [Zamani et al., 2023] and embodied in Taylor's stages of expressing an information need [Taylor, 1968]: Visceral, Conscious, Formalised, and Compromised. This proceeded into a discussion around how queries can be elicited from users and whether contextual cues can be used to improve the process either directly with the users or behind the scenes. Oard's idea of "querying by babbling" [Oard, 2012] was discussed as a potential mechanism to ease the burden while using generative systems to determine a query from a user's "babbling".

Related to this was a discussion of how we might better leverage the implied feedback in interactions with generative systems (e.g., "*no, that's wrong. Do this instead.*"). But it was expressed that this type of feedback might be hard to utilise effectively in the past [Bi et al., 2019]. Whether newer generative systems are better able to leverage this form of feedback is an interesting avenue to explore, especially if it can be combined with other form of relevance feedback to improve system results.

Bridging across all of these ideas was a discussion around how conversational interfaces may improve user interactions over search bars but still do not quite meet the group's needs. For example, to our knowledge no systems exist that can help tackle problems like "Find a flight from Sheffield to Washington under as cheaply as possible and no more than 10 hours." Such queries are only partially supported in even explicit flight searches now via advanced search options despite these types of criteria being frequent (at least for us). Moreover, the criteria may not always be explicit and may require other contextual clues given by past search history (e.g., "I have $500 decorate my room" but the room is not specified in the query).

On the other hand, such systems may also have to support different use cases simultaneously. This is not uncommon for e-commerce websites that have brick and mortar stores as well. There will be the general searching to buy something queries (e.g., bookshelves, a couch) but then there may also be specific queries to help locate an item in-store (e.g., finding what aisle/area Kallax shelves are stocked). Distinguishing between these types of queries and how to appropriately respond is worth exploring as they may be better addressed through different modalities.

The topic that crossed over all of this was a general disdain for the need to conduct "prompt engineering" and the idea that end-users would willingly and regularly engage in tailoring their queries to systems to optimise outcomes. While it may be feasible to assume this happens in a professional setting, we did not find a compelling time that this might occur for an every day information need.

Finally, we has a brief discussion around how generative systems may be incorporated into different forms of tasks whether it is for decision making processes, helping to complete tasks [Trippas et al., 2024], or managing personal information. While these are all compelling avenues to explore, they warranted more discussion than time allowed.

3.4 IR and Society

The IR community should reflect on the values and incentives encoded in our research—from what research is funded to what kind of impact we value collectively as a community. Our research does not exist in a void, rather it is deeply informed and shaped by the societal and organizational context in which they are embedded and conducted. Therefore, re-centering IR research on societal needs requires us to interrogate both the incentive and power structures that influence what IR research is done and what research we should be doing that is under-invested. This includes critically reflecting on the influence of big for-profit tech platforms on our research agenda and the corresponding concentration of power to make societally consequential decisions in the hands of these platform owners. It also motivates us to consider what research directions align with "IR for social good" that is being overlooked in the context of "IR for profit". Fundamentally, this begs the question of whose sociotechnical visions get to influence our community and whose perspectives are erased [Mitra, 2024]. What would IR research agendas look like if we broadened our focus from serving individual users with relevant content to focusing on critical needs of society such as knowledge production, health education, and promoting informed citizenry in democratic societies? How do we as a community make space for a multitude of radically different sociotechnical imaginaries to motivate our research and explicate the values guiding our own community?

There are several steps that we can imagine taking to realign IR research with societal needs. One such proposal may be to break-out of the reactionary loop where new emerging technologies trigger a flurry of critical research on their potential social consequences only to be interrupted by the next emerging breakthrough restarting the cycle. How can social good inform and motivate new IR research from the onset instead of being an afterthought of emerging new information access methods? How do we set our research agendas if our goal is to affect social good? A necessary component for this is to incorporate expertise from disciplines outside of IR that are critical to evaluating social impact. This motivates the need for spaces for interdisciplinary engagements and nurturing collaborations with other civil society stakeholders such as policy makers, activists, and artists. What resources do we need to help IR researchers navigate such interdisciplinary spaces? Should we create workshops and reading lists specifically to promote cross-pollination of ideas between disciplines? How do we correspondingly encourage individual researchers and practitioners to deeply consider the societal impact of their work regardless of their specific subjects of study? Do we want to encourage academic practices such as including reflections on ethical and social considerations of our research in our publications? Should we re-prioritize IR research on topics such as (i) "search as learning" [Collins-Thompson et al., 2017] that focuses education and knowledge production, and (ii) federated IR systems (*e.g.*, [Granitzer et al., 2024]) that decentralizes control over our information access systems. These are some of the questions that we must grapple with as a research community to ensure our work and their outcomes are aligned with an emancipatory humanistic future for all.

4 Final Note

The *First Search Futures Workshop* provided an opportunity for the community to voice their concerns about the emerging technologies and the potential impacts on the field and society at large. Overall, there was a shared sense of optimism and hope – full of new (and old) paths and directions to study, research and explore. In sum, the future is bright, the future is search.

A Authors and Affiliations

This is a list of authors of this publication. Workshop organizers (listed alphabetically by last name):

- Leif Azzopardi (University of Strathclyde, UK).
- Charles L. A. Clarke (University of Waterloo, Canada).
- Paul Kantor (University of Wisconsin Madison, USA).
- Bhaskar Mitra (Microsoft Research, USA).
- Johanne R. Trippas (RMIT University, Australia).
- Zhaochun Ren (Leiden University, The Netherlands).

Participants and speakers who contributed to this publication as well (listed alphabetically by last name):

- Mohammad Aliannejadi (University of Amsterdam, Netherlands).
- Negar Arabzadeh (University of Waterloo, Canada).
- Raman Chandrasekar (Institute for Experiential AI, Northeastern University, USA).
- Maarten de Rijke (University of Amsterdam, Netherlands).
- Panagiotis Eustratiadis (University of Amsterdam, Netherlands).
- William Hersh (Oregon Health & Science University, USA).
- Jin Huang (University of Amsterdam, The Netherlands).
- Evangelos Kanoulas (University of Amsterdam, Netherlands).
- Jasmin Kareem (Eindhoven University of Technology, Netherlands).
- Yongkang Li (University of Amsterdam, Netherlands).
- Simon Lupart (University of Amsterdam, Netherlands).
- Kidist Amde Mekonnen (University of Amsterdam, Netherlands).
- Adam Roegiest (Zuva, Canada).
- Ian Soboroff (NIST, USA).
- Fabrizio Silvestri (Sapienza University of Rome, Italy).
- Suzan Verberne (Leiden University, The Netherlands).
- David Vos (University of Amsterdam, The Netherlands).
- Eugene Yang (Johns Hopkins University, USA).
- Yuyue Zhao (University of Amsterdam, The Netherlands).

Acknowledgments

The organisers would like to thank all the speakers and participants for their valuable inputs, insights and feedbacks. And, we would also like to thank all the organisers of ECIR 2024 for hosting our workshop and for creating a very memorable and positive conference experience!

References

Syed Wajid Aalam, Abdul Basit Ahanger, Muzafar Rasool Bhat, and Assif Assad. Evaluation of fairness in recommender systems: A review. In *International Conference on Emerging Technologies in Computer Engineering*, pages 456–465. Springer, 2022.

Amin Abolghasemi, Suzan Verberne, Arian Askari, and Leif Azzopardi. Retrievability Bias Estimation Using Synthetically Generated Queries. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3712–3716, 2023.

Amin Abolghasemi, Leif Azzopardi, Arian Askari, Maarten de Rijke, and Suzan Verberne. Measuring Bias in a Ranked List Using Term-Based Representations. In *European Conference on Information Retrieval*, pages 3–19. Springer, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. Generative information retrieval evaluation. *arXiv preprint arXiv:2404.08137*, 2024.

Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *CHIIR*, pages 33–42. ACM, 2020.

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

Avishek Anand, Procheta Sen, Sourav Saha, Manisha Verma, and Mandar Mitra. Explainable information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3448–3451. Association for Computing Machinery, 2023.

Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *WWW*, pages 2155–2165, 2020.

Negar Arabzadeh and Charles LA Clarke. A comparison of methods for evaluating generative ir. *arXiv preprint arXiv:2404.04044*, 2024a.

Negar Arabzadeh and Charles LA Clarke. Fréchet distance for offline evaluation of information retrieval systems with sparse labels. *arXiv preprint arXiv:2401.17543*, 2024b.

Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. Adapting standard retrieval benchmarks to evaluate generated answers. In *European Conference on Information Retrieval*, pages 399–414. Springer, 2024.

Arian Askari, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, and Suzan Verberne. CLosER: Conversational Legal Longformer with Expertise-Aware Passage Response Ranker for Long Contexts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 25–35, 2023a.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5311–5315, 2023b.

Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, and Suzan Verberne. Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10087–10099, 2023c.

Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. Humanely: Human evaluation of llm yield, using a novel web-based evaluation tool. *medRxiv*, 2023. doi: 10.1101/2023.12.22.23300458. URL https://www.medrxiv.org/content/early/2023/12/30/2023.12.22.23300458.

Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D. Smucker. Report on the sigir 2010 workshop on the simulation of interaction. *SIGIR Forum*, 44(2):35–47, jan 2011.

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

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, et al. Megawika: Millions of reports and their sources across 50 diverse languages. *arXiv preprint arXiv:2307.07049*, 2023.

J.R. Baron, R.C. Losey, and M.D. Berman. *Perspectives on Predictive Coding: And Other Advanced Search Methods for the Legal Practitioner*. American Bar Association, Section of Litigation, 2016. ISBN 9781634256582. URL https://books.google.com/books?id=TdJ2AQAACAAJ.

Nicholas J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, jun 2008. ISSN 0163-5840. doi: 10.1145/1394251.1394261. URL https://doi.org/10.1145/1394251.1394261.

NJ Belkin and SE Robertson. Some ethical and political implications of theoretical research in information science. In *Proceedings of the ASIS Annual Meeting*, 1976.

Lucas Bernardi, Sakshi Batra, and Cintia Alicia Bruscantini. Simulations in recommender systems: An industry perspective. *arXiv preprint arXiv:2109.06723*, 2021.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen tau Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. *arXiv preprint arXiv:2204.10628*, 2022.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657. ACM, 2020.

Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. Conversational product search based on negative feedback. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 359–368, 2019.

Andrea Brennen. What do people really want when they say they want "Explainable AI?" We asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–7, New York, NY, USA, 2020. Association for Computing Machinery.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*, 2021.

Raman Chandrasekar, Harr Chen, Simon Corston-Oliver, and Eric Brill. Subwebs for specialized search. In *Proceedings SIGIR '04: Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 480–481, 2004.

Raman Chandrasekar, Dean A. Slawson, and Michael K. Forney. Temporal ranking of search results, 2006. US Patent US7849079B2 Filed 2006, Granted 2010.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *Proceedings of the 31st ACM International Conference on Information And Knowledge Management*, CIKM '22. ACM, 2022.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Continual learning for generative retrieval over dynamic corpora. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23. ACM, 2023a.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In *SIGIR 2023: 46th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1448–1457. ACM, July 2023b.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a reinforce recommender system. In *WSDM*, pages 456–464, 2019a.

Shi-Yong Chen, Yang Yu, Qing Da, Jun Tan, Hai-Kuan Huang, and Hai-Hong Tang. Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1187–1196, New York, NY, USA, 2018. Association for Computing Machinery.

Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. Generative adversarial user model for reinforcement learning based recommendation system. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1052–1061, 2019b.

Xiong-Hui Chen, Bowei He, Yang Yu, Qingyang Li, Zhiwei Qin, Wenjie Shang, Jieping Ye, and Chen Ma. Sim2Rec: A simulator-based decision-making approach to optimize real-world long-term user engagement in sequential recommender systems. *arXiv preprint arXiv:2305.04832*, 2023c.

Xuanang Chen, Ben He, Zheng Ye, Le Sun, and Yingfei Sun. Towards imperceptible document manipulations against neural ranking models. In *Association for Computational Linguistics: ACL*, 2023d.

Coalition for Health AI. Blueprint for Trustworthy AI Implementation Guidance and Assurance for Healthcare. Technical report, April 2023. URL https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf.

Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. Search as learning (dagstuhl seminar 17092). 2017.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. *ACM SIGIR*, 2024.

J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM New York, NY, USA, 2018.

Michiel de Jong, Yury Zemlyanskiy, Joshua Ainslie, Nicholas FitzGerald, Sumit Sanghai, Fei Sha, and William Cohen. Fido: Fusion-in-decoder optimized for stronger performance and faster inference. *arXiv preprint arXiv:2212.08153*, 2023.

Romain Deffayet, Thibaut Thonet, Jean-Michel Renders, and Maarten de Rijke. Offline evaluation for reinforcement learning-based recommendation: A critical issue and some alternatives. *SIGIR Forum*, 56(2), 2022.

Romain Deffayet, Thibaut Thonet, Dongyoon Hwang, Vassilissa Lehoux, Jean-Michel Renders, and Maarten de Rijke. SARDINE: A simulator for automated recommendation in dynamic and interactive environments. *arXiv preprint arXiv:2311.16586*, November 2023a.

Romain Deffayet, Thibaut Thonet, Jean-Michel Renders, and Maarten de Rijke. Generative slate recommendation with reinforcement learning. In *WSDM*, pages 580–588. ACM, 2023b.

Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, pages 1–50, 2023.

Dario Di Palma. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 1369–1373. Association for Computing Machinery, 2023.

Laura Dietz and John Foley. TREC CAR Y3: Complex answer retrieval overview. In *Proceedings of Text REtrieval Conference (TREC)*, 2019.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. TREC complex answer retrieval overview. In *Proceedings of Text REtrieval Conference (TREC)*, 2017.

Laura Dietz, Ben Gamari, Jeff Dalton, and Nick Craswell. TREC complex answer retrieval overview. In *Proceedings of Text REtrieval Conference (TREC)*, 2018.

V. Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, 2019. ISBN 9783030303716.

Karlijn Dinnissen and Christine Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data*, 5:913608, 2022.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*, 2021.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50, 2023.

Naghmeh Farzi and Laura Dietz. An exam-based evaluation approach beyond traditional relevance judgments, 2024.

Siyong Fu. A reinforcement learning-based smart educational environment for higher education. *Int. J. e-Collab.*, 19(6):1–17, dec 2022.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S^3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023.

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. Evaluating Generative Ad Hoc Information Retrieval, November 2023. URL http://arxiv.org/abs/2311.04694. arXiv:2311.04694 [cs].

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

Rachel S. Goodman, J. Randall Patrinely, Cosby A. Stone, Eli Zimmerman, Rebecca R. Donald, Sam S. Chang, Sean T. Berkowitz, Avni P. Finn, Eiman Jahangir, Elizabeth A. Scoville, Tyler S. Reese, Debra L. Friedman, Julie A. Bastarache, Yuri F. van der Heijden, Jordan J. Wright, Fei Ye, Nicholas Carter, Matthew R. Alexander, Jennifer H. Choe, Cody A. Chastain, John A. Zic, Sara N. Horst, Isik Turker, Rajiv Agarwal, Evan Osmundson, Kamran Idrees, Colleen M. Kiernan, Chandrasekhar Padmanabhan, Christina E. Bailey, Cameron E. Schlegel, Lola B. Chambless, Michael K. Gibson, Travis J. Osterman, Lee E. Wheless, and Douglas B. Johnson. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA network open*, 6(10):e2336483, October 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.36483.

Michael Granitzer, Stefan Voigt, Noor Afshan Fathima, Martin Golasowski, Christian Guetl, Tobias Hecking, Gijs Hendriksen, Djoerd Hiemstra, Jan Martinovič, Jelena Mitrović, et al. Impact and development of an open web index for open web search. *Journal of the Association for Information Science and Technology*, 75(5):512–520, 2024.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *CIKM*, 2016.

Shashank Gupta, Harrie Oosterhuis, and Maarten de Rijke. Safe deployment for counterfactual learning to rank with exposure-based risk minimization. In *SIGIR*. ACM, 2023.

Nam Le Hai, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. Cosplade: Contextualizing splade for conversational information retrieval. *arXiv preprint arXiv:2301.04413*, 2023.

Karyn Harty. Discovery program. In *Law Society Gazette*, volume 111, pages 44–47. Dublin, Ireland, April 2017.

Marti Hearst. User interfaces for search. *Modern Information Retrieval*, pages 21–55, 2011.

Natali Helberger. On the democratic role of news recommenders. *Digital Journalism*, 7(8):993–1012, 2019.

William Hersh. Search still matters: information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association: JAMIA*, page ocae014, January 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae014.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *KDD*, pages 1849–1858, 2015.

Ashley M. Hopkins, Jessica M. Logan, Ganessan Kichenadasse, and Michael J. Sorich. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI cancer spectrum*, 7(2):pkad010, March 2023. ISSN 2515-5091. doi: 10.1093/jncics/pkad010.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. *arXiv preprint arXiv:2305.08845*, 2023.

Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 190–199, 2020.

Jin Huang, Harrie Oosterhuis, Bunyamin Cetinkaya, Thijs Rood, and Maarten de Rijke. State encoders in reinforcement learning for recommendation: A reproducibility study. In *SIGIR 2022: 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2738–2748. ACM, July 2022a.

Jin Huang, Harrie Oosterhuis, and Maarten de Rijke. It is different when items are older: Debiasing recommendations when selection bias and user preferences are dynamic. In *WSDM 2022: The Fifteenth International Conference on Web Search and Data Mining*, pages 381–289. ACM, February 2022b.

Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. Retrieving supporting evidence for generative question answering. *arXiv preprint arXiv:2309.11392*, 2023.

Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. RecSim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847*, 2019a.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, et al. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. *arXiv preprint arXiv:1905.12767*, 2019b.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2021.

Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2022.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Moksh Jain and Sowmya Kamath S. Proximal policy optimization for improved convergence in irgan, 2019.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. *arXiv preprint arXiv:2301.01820*, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jiepu Jiang, Daqing He, and James Allan. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 607–616, 2014.

Karen Sparck Jones and Peter Willett. *Readings in information retrieval*. Morgan Kaufmann, 1997.

Evangelos Kanoulas, Mark Hall, Paul Clough, Ben Carterette, and Mark Sanderson. Overview of the trec 2010 session track. 2010.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14. Association for Computing Machinery, 2020.

Varsha Kishore, Chao Wan, Justin Lovelace, Yoav Artzi, and Kilian Q. Weinberger. Incdsi: Incrementally updatable document retrieval. *arXiv preprint arXiv:2307.10323*, 2023.

Vaibhav Kumar and Jamie Callan. Making information seeking easier: An improved pipeline for conversational search. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP '20*, pages 3971–3980. Association for Computational Linguistics, 2020.

F. Wilfrid Lancaster. *Information retrieval systems: Characteristics, testing, and evaluation*. John Wiley & Sons, New York, 2nd ed edition edition, January 1979. ISBN 978-0-471-04673-8.

Sunkyung Lee, Minjin Choi, and Jongwuk Lee. GLEN: Generative retrieval via lexical index learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7693–7704, Singapore, 2023. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval. *arXiv preprint arXiv:2306.15222*, 2023a.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Multiview identifiers enhanced generative retrieval. *arXiv preprint arXiv:2305.16675*, 2023b.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023c.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. Leveraging large language models for NLG evaluation: A survey. *arXiv preprint arXiv:2401.07103*, 2024.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *arXiv preprint arXiv:2005.02230*, 2021.

Dong Liu and Chenyang Yang. A deep reinforcement learning approach to proactive content pushing and recommendation for mobile users. *IEEE Access*, 7:83120–83136, 2019.

Feng Liu, Huifeng Guo, Xutao Li, Ruiming Tang, Yunming Ye, and Xiuqiang He. End-to-end deep reinforcement learning based recommendation with supervised embedding. In *WSDM*, pages 384–392. ACM, 2020.

Jingjing Liu and Nicholas J Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 26–33, 2010.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023a.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *International Conference on Information and Knowledge Management, CIKM*, 2023b.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Topic-oriented adversarial attacks against black-box neural ranking models. In *Conference on Research and Development in Information Retrieval, SIGIR*, 2023c.

Felicia Loecherbach, Judith Moeller, Damian Trilling, and Wouter van Atteveldt. The unified framework of media diversity: A systematic literature review. *Digital Journalism*, 8(5):605–642, 2020.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. Off-policy learning in two-stage recommender systems. In *WWW*, pages 463–473. ACM / IW3C2, 2020.

Shengjie Ma, Chong Chen, Qi Chu, and Jiaxin Mao. Leveraging large language models for relevance judgments in legal case retrieval. *arXiv preprint arXiv:2403.18405*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

David Maxwell and Leif Azzopardi. Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 1141–1144. ACM, 2016. doi: 10.1145/2911451.2911469. URL https://doi.org/10.1145/2911451.2911469.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. doi: 10.1145/3626772.3657846.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. *CoRR*, abs/2005.00661, 2020. URL https://arxiv.org/abs/2005.00661.

Yusuf Mehdi. Bringing the full power of copilot to more people and businesses, 2024. URL https://blogs.microsoft.com/blog/2024/01/15/bringing-the-full-power-of-copilot-to-more-people-and-businesses/.

Sanket Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. DSI++: Updating transformer memory with new documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8198–8213, Singapore, December 2023a. Association for Computational Linguistics.

Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. Dsi++: Updating transformer memory with new documents. *arXiv preprint arXiv:2212.09744*, 2023b.

Donald Metzler, Yi Tay, and Dara Bahri. Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum*, 55(1), 2021.

Bhaskar Mitra. Search and society: Reimagining information access for radical futures. *arXiv preprint arXiv:2403.17901*, 2024.

Ali Montazeralghaem, Hamed Zamani, and James Allan. A reinforcement learning framework for relevance feedback. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 59–68, New York, NY, USA, 2020. Association for Computing Machinery.

Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory and Discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T. Robertson, and Rasmus Kleis Nielsen. Reuters Institute digital news report 2023. Reuters Institute for the Study of Journalism, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf.

NIST. Artifical intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, NIST, January 2023.

Rodrigo Nogueira and Kyunghyun Cho. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 574–583, 2017.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine, November 2023. URL http://arxiv.org/abs/2311.16452. arXiv:2311.16452 [cs].

Douglas W. Oard. Query by babbling: a research agenda. In *Proceedings of the First Workshop on Information and Knowledge Management for Developing Region*, IKM4DR '12, page 17–22, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450317184. doi: 10.1145/2389776.2389781. URL https://doi.org/10.1145/2389776.2389781.

Daan Odijk, Edgar Meij, Isaac Sijaranamual, and Maarten de Rijke. Dynamic query modeling for related content finding. In *SIGIR 2015: 38th international ACM SIGIR conference on Research and development in information retrieval*, pages 33–42. ACM, August 2015.

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

Harrie Oosterhuis and Maarten de Rijke. Unifying online and counterfactual learning to rank. In *WSDM 2021: 14th International Conference on Web Search and Data Mining*. ACM, March 2021.

OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 11 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation, 2023.

Eli Pariser. *The Filter Bubble: What the Internet is Hiding from You*. penguin UK, 2011.

Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2022.

Ronak Pradeep, Kai Hui, Jai Gupta, Adam D. Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q. Tran. How does generative retrieval scale to millions of passages? *arXiv preprint arXiv:2305.11841*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. *CoRR*, abs/2104.06683, 2021. URL https://arxiv.org/abs/2104.06683.

Melrose Roderick, James MacGlashan, and Stefanie Tellex. Implementing the deep q-network. *arXiv preprint arXiv:1711.07478*, 2017.

Wilbert Samuel Rossi, Jan Willem Polderman, and Paolo Frasca. The closed loop between opinion formation and personalised recommendations. *IEEE Transactions on Control of Network Systems*, 9 (3):1092–1103, 2021.

Tetsuya Sakai. Swan: A generic framework for auditing textual conversational systems, 2023.

Rodrygo LT Santos, Craig Macdonald, Iadh Ounis, et al. Search result diversification. *Foundations and Trends® in Information Retrieval*, 9(1):1–90, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Chirag Shah and Emily M. Bender. Situating search. In *CHIIR*, pages 221–232. ACM, 2022.

Chirag Shah and Emily M. Bender. Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web*, 18(3), April 2024.

Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and An-Xiang Zeng. Virtual-Taobao: Virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4902–4909, 2019.

Herbert A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482, 1962.

Dean A. Slawson, Raman Chandrasekar, and Michael K. Forney. Adaptive dissemination of personalized and contextually relevant information, 2006. US Patent US7577718B2 Filed 2006, Granted 2009.

EuiYul Song, Sangryul Kim, Haeju Lee, Joonkee Kim, and James Thorne. Re3val: Reinforced and reranked generative retrieval. *arXiv preprint arXiv:2401.16979*, 2024.

Natalie Jomini Stroud. *Niche News: The Politics of News Choice*. Oxford University Press, 2011.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. Learning to tokenize for generative retrieval. In *NeurIPS 2023: Thirty-seventh Conference on Neural Information Processing Systems*, December 2023.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR*, 2014.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. Semantic-enhanced differentiable search index inspired by learning strategies. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 4904–4913, New York, NY, USA, 2023a. Association for Computing Machinery.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. Recent advances in generative information retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 294–297, 2023b.

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

Yi Tay, Vinh Quang Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In *NeurIPS*, 2022.

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

Maartje Ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, Ron Mulder, Nick van der Wildt, and Maarten de Rijke. Do news consumers want explanations for personalized news rankings? In *FATREC Workshop on Responsible Recommendation*, August 2017.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

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 *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, New York, NY, USA, 2024. ACM. doi: 10.1145/3626772.3657914.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards Conversational Diagnostic AI, January 2024. URL http://arxiv.org/abs/2401.05654. arXiv:2401.05654 [cs].

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 355–363, New York, NY, USA, 2021. ACM. ISBN 9781450382977.

Liesbet Van Bulck and Philip Moons. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value and danger of ChatGPT-generated responses to health questions. *European Journal of Cardiovascular Nursing*, page zvad038, April 2023. ISSN 1873-1953. doi: 10.1093/eurjcn/zvad038.

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *ACM SIGMOD Conference*, 2004.

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 173–183, 2021.

Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten de Rijke. RADio–Rank-aware divergence metrics to measure normative diversity in news recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 208–219, 2022.

Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55, 2010.

William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1):14045, September 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-41032-5.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023a.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Allen Sun, Weiwei Deng, Qi Zhang, and Mao Yang. A neural corpus indexer for document retrieval. *arXiv preprint arXiv:2206.02743*, 2023b.

Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. Novo: Learnable and interpretable document identifiers for model-based ir. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 2656–2665, New York, NY, USA, 2023c. Association for Computing Machinery.

Tom Warren. Microsoft's new copilot pro brings ai-powered office features to the rest of us, 2024. URL https://www.theverge.com/2024/1/15/24038711/microsoft-copilot-pro-office-ai-apps.

Zeng Wei, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. Reinforcement learning to rank with markov decision process. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 945–948, New York, NY, USA, 2017. Association for Computing Machinery.

Ryen W. White. Tasks, copilots, and the future of search. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. PRADA: practical black-box adversarial attacks against neural ranking models. *ACM Trans. Inf. Syst.*, 41(4): 89:1–89:27, 2023.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E. Ho, and James Zou. How well do LLMs cite relevant medical references? An evaluation framework and analyses, February 2024. URL http://arxiv.org/abs/2402.02008. arXiv:2402.02008 [cs].

Jun Xu, Zeng Wei, Long Xia, Yanyan Lan, Dawei Yin, Xueqi Cheng, and Ji-Rong Wen. Reinforcement learning to rank with pairwise policy gradient. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 509–518, New York, NY, USA, 2020. Association for Computing Machinery.

Eugene Yang, David D. Lewis, and Ophir Frieder. On minimizing cost in legal document review workflows. In *Proceedings of the 21st ACM Symposium on Document Engineering*, 2021.

Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. Auto search indexer for end-to-end document retrieval. *arXiv preprint arXiv:2310.12455*, 2023.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. KG-FiD: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4974, Dublin, Ireland, 2022. Association for Computational Linguistics.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. Few-shot generative conversational query rewriting. *arXiv preprint arXiv:2006.05009*, 2020.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM, 2021.

Yifei Yuan, Clemencia Siro, Mohammad Aliannejadi, Maarten de Rijke, and Wai Lam. Asking multimodal clarifying questions in mixed-initiative conversational search. In *The Web Conference 2024*. ACM, May 2024.

Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. Retrieval-enhanced machine learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2875–2886, 2022.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456, 2023.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part I: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36, 2022a.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Computing Surveys*, 55(6):1–41, 2022b.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. Scalable and effective generative information retrieval. *arXiv preprint arXiv:2311.09134*, 2023.

Wei Zeng, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. Multi page search with reinforcement learning to rank. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, page 175–178, New York, NY, USA, 2018. Association for Computing Machinery.

Saber Zerhoudi, Sebastian Günther, Kim Plassmeier, Timo Borst, Christin Seifert, Matthias Hagen, and Michael Granitzer. The simiir 2.0 framework: User types, markov model-based interaction simulation, and advanced query generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 4661–4666. ACM, 2022. doi: 10.1145/3511808.3557711. URL https://doi.org/10.1145/3511808.3557711.

Hailin Zhang, Yujing Wang, Qi Chen, Ruiheng Chang, Ting Zhang, Ziming Miao, Yingyan Hou, Yang Ding, Xupeng Miao, Haonan Wang, Bochen Pang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, Xing Xie, Mao Yang, and Bin Cui. Model-enhanced vector index. *arXiv preprint arXiv:2309.13335*, 2023a.

Peitian Zhang, Zheng Liu, Yujia Zhou, Zhicheng Dou, and Zhao Cao. Term-sets can be strong document identifiers for auto-regressive search engines. *arXiv preprint arXiv:2305.13859*, 2023b.

Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020.

Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. Recommendations with negative feedback via pairwise deep reinforcement learning. In *KDD*, pages 1040–1048. ACM, 2018.

Xiangyu Zhao, Xudong Zheng, Xiwang Yang, Xiaobing Liu, and Jiliang Tang. Jointly learning to recommend and advertise. In *KDD*, pages 3319–3327. ACM, 2020.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In *WWW*, pages 167–176. ACM, 2018.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2023.

Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, and Ji-Rong Wen. Dynamicretriever: A pre-training model-based ir system with neither sparse nor dense index. *arXiv preprint arXiv:2203.00537*, 2022a.

Yujia Zhou, Jing Yao, Zhicheng Dou, Ledell Wu, Peitian Zhang, and Ji-Rong Wen. Ultron: An ultimate retriever on corpus with a model-based indexer. *arXiv preprint arXiv:2208.09257*, 2022b.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. Enhancing generative retrieval with reinforcement learning from relevance feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12481–12490, Singapore, 2023. Association for Computational Linguistics.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128*, 2023.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. Large language models are built-in autoregressive search engines. *arXiv preprint arXiv:2305.09612*, 2023.

Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. Reinforcement learning to optimize long-term user engagement in recommender systems. In *KDD*, pages 2810–2818, 2019.