Report on the Workshop on Task Focused IR in the Era of Generative AI

Chirag Shah
University of Washington
Seattle, WA, USA
chirags@uw.edu

Ryen W. White
Microsoft Research
Redmond, WA, USA
ryenw@microsoft.com

Abstract

Search and recommender systems should prioritize support for user tasks over support for individual queries or actions. To that end, the Information Retrieval (IR) community has spent decades trying to understand users' tasks and their contexts, and how to best assist users in making progress toward completing them. The recent advancements in generative artificial intelligence (AI) have drastically shifted the landscape of task-focused IR. Users can now express not only their queries and questions, but actual information needs, tasks, and goals in natural language to an AI system and receive not just results, but also answers in natural language that are generated specifically for them. This new paradigm raises many interesting questions, opportunities, and challenges. We brought together a group of highly motivated students and scholars in a two-day workshop on the Microsoft campus in Redmond to discuss these issues, learn from each other, and envision a new future for task-focused IR and information access more broadly.

Date: September 28-29, 2023.

Website: https://ir-ai.github.io.

1 Introduction

We originally envisioned a workshop focused on task-based IR. That was in 2020 and while we had a lot of interest and funding from US National Science Foundation (NSF) and from Microsoft, the COVID-19 pandemic upended those plans. For almost three years we waited for the right time and a good opportunity to organize this event. By mid-2023, we were convinced that we had that in spades. The pandemic was behind us and generative AI (GenAI) had disrupted many areas in IR. Specifically, foundation models (GPT, DALL·E, etc.) were starting to change how we understand and address tasks in IR. We always wanted to organize this workshop at Microsoft in Redmond over one-and-a-half days (Thursday-Friday). We felt that spanning multiple days instead of fitting everything into one day would allow the participants to socialize and connect more, remove some time pressure, and allow us to adapt the program dynamically, depending on the flow the event and feedback from participants about where they wanted to spend time. In addition, we also felt that ending the workshop on a Friday with a lunch would allow the participants to either travel back home that day or choose to spend the weekend in the Seattle area.

The workshop was organized in a whirlwind of activity over a short period of time. Once we identified the dates (September 28-29, 2023), we quickly created a website with a tentative plan and a call for participation, and started advertising. The funds from NSF would enable us to cover travel expenses for those US-based students attending the workshop from outside of the Seattle area, but not anyone else. Thanks to support from ACM SIGIR (through the "Friends of SIGIR" program), we could cover student travel costs for students coming from Canada. Funds from The Information School (iSchool) at University of Washington were used for catering. Finally, Microsoft provided the venue (a lecture room in Building 99, the home of Microsoft Research in Redmond), registration support, and local logistics.

The timeline for registration and travel planning was very tight – just about a month between the announcement and the workshop. The students were also given an opportunity to submit an abstract for their posters. All of these activities – poster submission, notifications, and registration – happened in the first two weeks of September 2023. Despite these constraints, we were fortunate to have a huge interest in this workshop. In the end, we had nearly 50 participants attending in person. Two third of these were students, one fifth were faculty, and the rest were industry professionals. From the students, two thirds were in a PhD program and the rest were in a master's program. Most participants were from the US, with about half a dozen traveling from Canada.

We received several requests to join the workshop virtually, but due to logistical issues and a very tight timeline, we could not make this a hybrid event. We did record almost everything except the breakout sessions and have made presentation slides and video recordings available for the community on the workshop website (see the hyperlink above).

2 Workshop Program

The program for the workshop included a mixture of talks, a panel, a poster session (and accompanying lightning talks), and two breakout group discussions, one on each day. Since the breakouts intentionally occupied a lot of the time at the workshop, we focus on them in Sections 3 and 4.

Nick Craswell from Microsoft gave a keynote on "Personalization and Conversation (and GPT-4)." In addition, we had three invited short talks, given by Luanne Sinnamon (University of British Columbia), Grace Hui Yang (Georgetown University), and Raman Chandrasekar (Northeastern University). On Day 1 of the workshop, we also organized a panel moderated by Rob Capra (University of North Carolina at Chapel Hill) with panelists Leif Azzopardi (University of Strathclyde), Jacek Gwizdka (University of Texas at Austin), and Besmira Nushi (Microsoft Research). The poster session on Day 1 was designed for the students to showcase their work and get feedback. Ten students from different institutions across the US and Canada participated in the poster session. They first gave a lightning talk and then presented their posters in an interactive format. In addition, the coffee and lunch breaks served as catalysts for continuing some of the conversations from posters and breakouts, and forming connections for potential collaborations.

3 Themes from Breakout Session 1

On Day 1 of the workshop, participants self-selected into breakout groups, each with a pre-selected senior participant playing the role of "captain" to help keep the discussions on track. Each group was assigned three questions. The questions and themes in their responses are summarized below.

1. What role(s) can GenAI play in addressing IR problems?

Emergent themes from participants in the breakout groups included:

- *Intent Understanding:* GenAI can clarify users' information needs and help them express their needs, especially for exploratory tasks. This can be done through open-ended dialog, reference interviews, and clarification questions. This also applies retrospectively to log analysis.
- *Query Support:* Given an original query, GenAI can perform query expansion and rewriting, especially when queries are not full sentences. It can also measure query characteristics (e.g., ambiguity, difficulty) and adjust/augment queries to provide more accurate results.
- *Corpus Understanding:* GenAI, acting as a reference librarian, can understand the collection and help users navigate through it.
- *Synthesizing Results:* GenAI can synthesize results from one or more queries, improving the efficiency of information extraction.
- *Ranking and Re-ranking:* GenAI can assist in ranking and re-ranking search results based on user input, enhancing the relevance of the results shown to users.
- *Evidence Attribution:* GenAI can connect answers to sources. Evidence attribution can solve the problem of evaluating accuracy or correctness of the GenAI output. It helps build a scaffold to get to the correct answer.
- *Human in the Loop:* GenAI can work with human input to improve the retrieval model and provide more personalized results.
- *Evaluation:* GenAI can generate labels for evaluation and create simulations that enable more diversity and realism in user models. It can establish criteria for the evaluation of its own answers (or those from other models) and use that feedback to self-improve. GenAI can also triage and reply to qualitative responses and evaluate them.

In summary, participants felt that GenAI has a lot of applications at different stages of the search process, from understanding intents to surfacing results and direct answers. Participants also mentioned applications of GenAI thereafter too, e.g., evaluating system responses, using feedback to improve search quality, and analyzing retrospective log data collected during system use.

2. Where should we not use GenAI tools like LLMs for IR problems? Why?

Examples of themes from participants in the breakout groups included:

- *Safety Concerns:* GenAI tools have the potential to be dangerous in some situations, especially when they may generate or propagate harmful, misleading, or biased information. This is a significant concern, especially when dealing with sensitive or critical topics.
- *High-Stakes Domains:* GenAI tools may not be suitable for high-stakes domains where the consequences of incorrect or incomplete information could be severe. This includes areas where people's health, safety, or legal matters are at stake.
- *Low Domain Knowledge:* Users with low domain knowledge may not be able to critically evaluate the information retrieved by GenAI tools. This can lead to misunderstandings, misinterpretations, or reliance on incorrect information.

- *Task Risk Variability:* Some tasks may appear low risk in certain situations but become high risk in others. GenAI tools should not be used without careful consideration of the specific context and potential risks associated with the task.
- *Intrusion:* GenAI should not intrude too much into the user's search process. They should follow mixed-initiative design guidelines and respect the need for human control over the interaction.
- *Implicit Influence:* GenAI should not be used to implicitly influence certain groups of users or tailor search behaviors, as this can involve human bias and affect people negatively.
- *Morality and Ethics:* GenAI should not be used in situations where moral and ethical decisions are required, as these decisions often involve complex human judgment that GenAI may not fully comprehend.
- *Precision Cases:* In cases where precise answers are required, GenAI responses might not be suitable since the answers they offer tend to be more general.
- *Replacing Human Assessors:* GenAI should not replace human gold standard assessors. Relevance is best understood by humans. AI can at best assist humans in determining relevance.
- *Reproducibility:* The outputs of GenAI models are non-deterministic. If the same GenAI cannot reproduce the same outputs over time, then the outputs may not be reliable.
- *Resource Costs:* The training and inference cost is a key consideration. The ecological and financial cost of using GenAI should be monitored and reduced where possible.

In summary, participants felt that while GenAI can be beneficial in many IR problems, it should not be used in situations where high precision, reproducibility, and accuracy are required, or where it could significantly impact people's lives. safety, accuracy, and domain knowledge are crucial intrude on the user's process, amplify biases, or make moral decisions. Furthermore, participants wanted the efficiency of GenAI models to be improved to cut costs and reduce latencies. They also believed that GenAI should also be used in applications where it augments human capabilities not replace humans, especially in contexts where human judgment and expertise are essential.

3. What are some of the problems created while addressing these challenges? Are there ways to address them effectively and still leverage the benefits of GenAI?
There are several problem themes highlighted by participants, for example:

- *Transparency:* Ensuring transparency in how GenAI operates and make decisions. This is crucial for building user trust and understanding.
- *Bias Amplification:* The use of GenAI can lead to misrepresentation of gender and other biases. GenAI can sometimes amplify users' implicit biases. The challenge is to make GenAI provide unbiased assistance. GenAI may leverage users' cognitive biases, reinforce the biases, and unintentionally manipulate users' judgments on information and decision-making. If GenAI learns from biased users, they could create echo chambers, reinforcing user biases.
- *Trustworthiness:* Ensuring the trustworthiness of the information, preventing hallucination (generation of information not present in the input data), and avoiding misinformation are critical in GenAI applications. There is a risk that GenAI could be used to generate misinformation, which needs to be mitigated. For topics such as vaccines, the freshness and temporal utility of the information are important, and GenAI struggles to answer on anything not in its training data. Using retrieval-augmented generation and citing sources can help address these issues.

- *Interaction and Control:* Users want to stay in control and understand why answers were provided by GenAI-based systems. The user experience plays an important role, and there are emerging opportunities for voice and other multimodal interactions.
- *Scalability:* GenAI models are large and expensive to train and deploy, which can be a barrier to their widespread use. Small language models are emerging that are specialized to certain tasks and applications, including search.
- *Data Sources:* Different sources for GenAI can lead to variations in the quality and reliability of the information generated. There are challenges in training data quality (and soon, also quantity) and in creating an information ecosystem, including access to third-party information, dealing with paywalls, and handling sponsored content and advertisements. Citing sources with links can help retain incentives for content creators (and advertisers/publishers) to participate.
- *Prompt Engineering:* There are challenges in designing prompts for GenAI-based systems, especially for non-AI experts. Lightweight training and tutorials can help users (crafting specific prompts expressing their intentions) and system designers (controlling system functionality).

Beyond the issues above, participants also raised other, more nuanced, issues, such as the challenges with achieving highly precise answers given the probabilistic nature of GenAI models and the potential to employ adversarial training to improve GenAI performance and reduce errors.

4 Themes from Breakout Session 2

On Day 2 of the workshop participants spent time in their breakout groups discussing how to build on the conversations and learnings in Day 1 and create some plans for meaningful outputs from the workshop. The outputs could take a variety of forms including perspective papers, research questions, research proposals, and policy recommendations, and span a variety of themes, summarized at a high level in the paragraphs below.

Interaction: Participants focused on comparing traditional search methods with interactions enhanced by GenAI. The discussion covered various aspects, including methods (e.g., queries vs. long prompts, length of interactions, number of iterations, task types, user expectations, and transferability of knowledge from search studies to GenAI scenarios), interface design (e.g., search interfaces for GenAI, the need to get creative given that short conversations with GenAI systems can be common), and evaluation metrics (e.g., system-side metrics such as the desire for more engagement, user-side metrics such as effectiveness and satisfaction). The session also provided some design recommendations, including promoting the use of GenAI as a collaborator to decompose complex tasks, and posed research questions on understanding the types of interactions with GenAI, the impact of GenAI on content consumption, and the impact of GenAI on search as a learning process. More broadly, questions were also raised on issues such as understanding past experiences with AI interaction, particularly in social media and among vulnerable populations, the shift in signals from clicks to chats in conversational interaction models, and how best to integrate GenAI with search interfaces.

Specialization: Participants debated a monolithic approach and a federated approach for GenAI in system design. The monolithic approach envisions a single, all-encompassing GenAI model (such as GPT-n) capable of addressing any problem, while the federated approach involves specialized models addressing specific tasks or subtasks. Flexibility was highlighted as a tradeoff, with the need to consider task-specific interfaces. The discussion emphasized that certain tasks should not solely depend on GenAI and that other sources of evidence are needed, especially for consequential tasks such as health.

Personalization: Participants also discussed adapting GenAI-based search experiences to individual users and the potential of GenAI to afford new ways of personalization by combining and synthesizing data to create personalized media feeds and learn from users' individual information interactions. They also discussed generating documents or information from data plugins to enable personalized content creation, addressing the challenge of tailoring system responses to individual users and their knowledge, and personalizing search experiences based on human mental models.

Evaluation: Multiple perspectives on evaluating GenAI in IR were discussed. The overall theme revolved around the need for comprehensive evaluation methods that consider traditional metrics (e.g., answer relevance), accuracy (e.g., verification of answers via external knowledge graphs, training GenAI models to recognize their own hallucinations), fluency (e.g., evaluate GenAI outputs using existing NLP methods, fluency scores, perplexity, grammar checkers, etc. to understand generation quality), human perceptions (e.g., multi-dimensional evaluation, with metrics such as completeness, alignment, trustworthiness, novelty, utility, etc.), and robustness to query variations (including adversarial perturbations) when assessing the performance of GenAI-based search systems. Beyond system evaluation, participants also discussed understanding the human impacts of GenAI, e.g., collecting new signals beyond IR for understanding human learning in GenAI-based search systems.

Feedback: The focus was on evaluating feedback mechanisms for content generated by GenAI. The discussion covered existing approaches used by systems such as Bing Chat, Bard, and ChatGPT (e.g., thumbs up, likes, flags, feedback links), potential issues (e.g., no incentives for users, no updates on issue resolution), and alternative approaches to designing user interfaces for gathering feedback to enhance engagement, including providing users with a record of their feedback, offering links to forum pages, a bug bounty program, and gamification and leaderboards.

Responsibility: Participants discussed the need to examine the role of GenAI and the responsibility of users and developers in using AI tools. They raised concerns about the dangers of using GenAI for consequential (e.g., health-related) queries, the potential for misinformation and bias in search results, and the implications of these risks for information interaction. They highlighted issues of bias, diversity, and hallucination in GenAI-based search systems, drew parallels with past experiences in human interaction with technology, and flagged a need to examine the effects of the size and shape of the model on issues such as hallucinations, fairness, and bias. Diving deep on diversity, participants highlighted the overrepresentation of English in mainstream GenAI models and its potential impact on the digital divide. They drew parallels with medical studies to underscore the importance of diversity in AI system design. Participants proposed a large-scale study to collect queries and relevance judgments from diverse communities to evaluate GenAI-based search systems comprehensively, including in various domains, such as coding, customer service, and marketing.

Policy and Regulation: Participants discussed the need for clear regulations and guidelines to ensure the responsible use of GenAI models. Regulation was mentioned, exploring questions about expected behaviors, differentiation based on user demographics, and the possibility of US federal regulations. The idea of high-level rules, akin to Asimov's laws of robotics, was also proposed, recognizing occasional rule breaking. The comparison with the regulation of corporations and the call for minimizing harm through regulations in social media and search engines were also discussed. The discussion extended to policy regulation, contrasting idealism and skepticism, viewing AI as a corporation, and considering the humanistic aspects of technology. Social media was scrutinized in discussions for its role in misinformation and the power disparities it might create. Overall, though, alignment, safety, and the need for rational views on GenAI were emphasized over regulation, with considerations about the alignment of GenAI "values" with human values. Participants also explored the spectrum of anthropomorphizing AI and its implications.

Ethics: The discussion also touched upon various ethical considerations such as GenAI systems learning from data, legal liability, explainability to stakeholders, transparency, privacy concerns, consequences of malicious use, potential lock-in or stickiness of users to a specific system, and user perception and awareness of GenAI.

Capabilities: Participants discussed how GenAI can be applied to various tasks in the domain of IR, including language-related tasks such as sentence completion, chatbot-style conversations, healthcare applications (e.g., providing medical information, offering emotional/therapy support), education/tutoring, entertainment, query expansion, question prediction, recommendations, information extraction, summarization, document expansion, code-related tasks (e.g., completion, generation, translation), image/video processing (e.g., medical imaging, deepfakes), and audio tasks (e.g., automated phone calls, music generation, text-to-speech).

Limitations: Participants also discussed some opportunities for GenAI in the domain of IR, including its current struggles with include long-form story generation, inference on out-of-domain content, cold start problems, handling complex math/logic, nuanced understanding, addressing diversity and inclusion, ensuring accessibility, considering risk factors, dealing with consequences of false positives/negatives, potential effects on third parties, handling paywalls, adapting to changes in the advertising ecosystem, addressing economic and job implications, avoiding anthropomorphization, managing efficiency and environmental impact, understanding two-way effects in user-system interactions, and navigating feedback cycles.

Future of Search: Finally, participants discussed if and how GenAI-based search differs from traditional IR. Participants sought to clarify what GenAI means for search, emphasizing that it encompasses more than just large language models but instead includes the full search experience. Participants discussed whether information will be retrieved or created based on individual preferences in the future. They also discussed potential disruptions to social media and media distribution models and moving beyond traditional documents to a broader range of sources for search. The importance of investigating larger information ecosystems, engaging stakeholders, and considering technology as humanity's offspring was also highlighted, as was the need to consider and develop ubiquitous multimodal IR systems that can serve relevant answers for any request in any format (including images, video, audio, etc.) at any time.

5 Post-Workshop Feedback

We solicited anonymous feedback from the workshop participants using a simple survey distributed after the event. The survey asked them about their experiences along different dimensions of the workshop and what could be improved for future events. The survey was completed by 18 participants (36% response rate), out of which 13 were students. Overall, the respondents found the workshop to be very useful (average usefulness rating of 4.2 out of 5, all but one rated 4+).

Participants very much appreciated the opportunity to meet in person with others interested in this space, having networking and collaboration opportunities, and learning from many experts in the field. One student summed it up, “Thank you very much for organizing this workshop! It was also great opportunity to build connection with other students. I was eager to this kind of opportunity, particularly as a student who started the PhD program during the pandemic.” The feedback received from other participants mirrored this sentiment.

6 Next Steps

Given the success of this event and the positive feedback that we received throughout the workshop and afterwards, we are now planning to do a follow-up event. That event may not be the same kind of workshop, but we know that we have a lot of interest and many people wanting to get together again and perhaps produce more tangible outcomes. In addition, we are currently considering ways to publish the collected works by the workshop participants (and others interested in this area). Some of these possibilities include a special issue of a journal and an edited book.

7 Acknowledgments

This workshop was made possible by generous support from the US National Science Foundation (NSF) award IIS-2023924, The Information School (iSchool) at the University of Washington, Microsoft Research, and ACM SIGIR.