Report on the 2nd 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

Generative Artificial Intelligence (GenAI) is revolutionizing how people access information and how they tackle and complete complex information tasks. This report is a summary of a recent workshop at Microsoft on this important and pressing topic. The event brought together a diverse mix of attendees from different professions and at different career stages for an engaging day of presentations and discussions. The emergent themes are described in detail in this summary.

Date: 27 September 2024.

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

1 Introduction

The second workshop on "Task Focused Information Retrieval in the Generative AI Era" was held on September 27, 2024, hosted by Microsoft Research on the Microsoft campus in Redmond, Washington. Around 60 participants from various organizations – academic and industrial – and various positions – students, faculty, professionals – from across the United States came together for this one day event on discussing issues related to information retrieval and access systems in the context of GenAI, specifically GenAI tools such as Large Language Models (LLMs). More information on the workshop, including the agenda, is available at https://ir-ai.github.io.

The workshop commenced with an inspiring keynote by Dr. Milad Shokouhi, Partner Group Science Manager at Microsoft. Dr. Shokouhi's presentation described a GenAI-based conversational assistant, the Microsoft 365 (M365) Copilot,[1] that provides contextual recommendations and responses to help enhance user productivity. The M365 copilot is integrated with other Microsoft applications, such as Teams, to help with tasks such as discovery, summarizing, drafting, and automating workflows. The keynote focused on how the copilot functions at a high-level, and some of the challenges and learnings from its development and deployment.

Following the well-received keynote, workshop participants were asked to come up with a set of specific questions or topics pertaining to the larger area of task-focused Information Retrieval (IR)

[1] https://www.microsoft.com/en-us/microsoft-365/copilot

systems in the context of GenAI. Dozens of questions, ideas, and topics were posted on a large whiteboard using sticky notes. Participants then arranged these notes into four broad categories: (1) theory, (2) benchmarks and evaluation, (3) users and user experience, and (4) applications and integration.

For the remainder of the day, we organized breakout sessions where the participants used the notes for the corresponding topics to stem their discussions and expand on their ideas. The groups took notes in a shared document. The following sections summarize the key points from their notes and the discussions.

2 Theory

While there were many threads of discussions on various theoretical constructs in GenAI, such as context, language, and interactions, the groups spent a significant amount of time talking about re-learning (updating the model knowledge or capabilities based on new data or feedback), unlearning (removing knowledge learned during training, e.g., for privacy, copyright, etc.), and readjustments for LLMs when it comes to information access. This is particularly needed to address issues of privacy, bias, and toxicity while also providing a more flexible architecture for further learning and refinements. For example, the following approaches were discussed for unlearning in LLMs:

1. **Training the Foundational Model**: This approach was found to be not feasible in most situations due to the need to retrain the model for every data removal request.
2. **Decoding Strategies**: This will involve preventing generation of certain tokens. However, models might find alternative ways to express similar intents.
3. **Guardrails/Censorship**: This idea requires implementing a layer to discourage certain topics and training the LLM to provide more diplomatic answers instead of deleting information.
4. **Reinforcement Learning from Human Feedback (RLHF)**: This popular technique to align LLMs with human preferences [Ziegler et al., 2019] can be used after initial training to discourage specific concepts/tokens.

Workshop participants also discussed alternative ways to train a foundational model for better, more flexible and nuanced training, e.g.,

1. **Speculative Decoding**: This approach is based on student-teacher concept with a small and large model where they decode tokens sequentially and a large model that verifies tokens as they go. This approach improves efficiency and has been found that it does not affect accuracy [Leviathan et al., 2023].
2. **Segmented Corpuses**: This approach involves training different segments of a large corpus based on expertise for specific motives.
3. **Multi-agent Auditing**: Use experts to prevent other LLMs from generating unlearned content.
4. **Distributed Models vs. Single/Centralized Model**: Mimic the human neuron system for more efficient inference and storage.
5. **Graph/Network of Models**: Each node is responsible for a specific concept, requiring sufficient common ground for communication.

3 Benchmarks and Evaluation

Two breakout groups focused on issues related to evaluation, datasets, and benchmarks for using GenAI for information access applications. The participants emphasized the importance of reliability and validity in evaluating and benchmarking LLMs. They noted that before establishing benchmarks, it is crucial to ensure that both the benchmarks and the LLMs themselves are reliable and valid. This foundation is necessary to address issues of fairness, bias, and equity.

The groups highlighted the need for shared definitions of key terms and discussed how metrics should evolve to be more meaningful within specific tasks. Benchmarks should be context-specific to provide accurate evaluations.

When it comes to business use cases and **personas**, the discussion focused on evaluating personalization effectiveness in relation to human preferences, laws, and values. The participants explored how to structure use cases, noting that product design often uses "personas" to capture diverse user needs. However, it is challenging to cover all user differences with benchmarks, leading to questions about grouping users and assessing personalization without creating echo chambers or experiencing distribution collapse.

The groups also addressed the need for data to perform reliable evaluations. They discussed the scarcity of comprehensive **open-source data** and suggested two solutions: using community data collection and encouraging organizations to release data collaboratively. Maintaining the quality of human evaluation was another key point. The participants emphasized the importance of context-specific questions to get accurate feedback.

On the topic of **alignments and ethics**, they discussed aligning safety and ethical principles, evaluating alignment success, and maintaining privacy.

Finally, the groups touched on the concept of **knowing**—specifically, how to get models to acknowledge when they do not know something. They suggested including confidence intervals in outputs and having models confirm or paraphrase inputs to improve transparency and reliability.

The discussion also covered the potential to teach LLMs appropriateness through system-level **content moderation**, including parental controls, flags, and guardrails. They considered the importance of reading levels and the classification and generation of documents, noting that higher volumes of content could bypass filters.

Evaluating **appropriateness** was another key topic. The group suggested using personalization algorithms to measure what is appropriate, understanding negative feedback, and utilizing both explicit and implicit user feedback to improve satisfaction. They also mentioned the importance of historical behavior logs and cultural evaluation and alignment, noting that standards change over time.

Context was highlighted as crucial, with understanding intent being particularly challenging due to fuzzy boundaries and user subjectivity. The ability to solve complex queries and provide feedback interfaces for improvement was also discussed, along with fine-tuning for pluralistic alignment.

Finally, the group discussed setting contextual measures and measuring controllability, emphasizing the need for dynamic and temporal evaluation and the ability of LLMs to evaluate higher-level constructs.

4 Users and User Experience

In the breakout groups for discussing users and user experience, participants delved into the intricacies of enhancing user interaction and trust in LLMs. They began by emphasizing the importance of referring to "people" instead of "users" to better capture the human aspect of these interactions.

The conversation then shifted to the **typology of tasks** that these people do, highlighting the need for systems that can effectively respond to various goals and intentions. For instance, assisting someone in learning how to apply for a green card requires a nuanced understanding of their needs, circumstances, and queries.

The usefulness of LLMs was discussed, with a focus on how it depends on both the individual and the system. Understanding the user involves considering the language used in queries, persona/user modeling, and cultural sensitivity. The group debated whether to curate pre-training data for users or to employ post-processing training methods.

Developing robust **user simulators** emerged as a critical point, as current interaction patterns with LLMs are not well-defined. The challenge lies in creating a "good enough" user simulator that accurately reflects real-world interactions.

Participants noted that while users may prefer simpler answers, which can increase the acceptance of LLMs, this preference can also lead to **misinformation**. Balancing user engagement with well-being is crucial.

Extending the issue of misinformation, the discussion steered towards **ethical considerations**. The group explored who controls the data, ownership, and access, questioning whether LLMs should always provide certifiable truths and discussing the broader social responsibilities of these models.

Building **trust** was identified as fundamental. It is essential for LLMs to acknowledge when they do not know something. Using prompts to eliminate out-of-bound questions and effectively conveying uncertainty were highlighted as vital strategies for building trust.

The group also debated the necessity of pseudo-relevance feedback versus using LLMs to formulate queries. They explored whether a new form of **relevance feedback**, more suited to the LLM era, is needed.

Designing systems that provide balanced perspectives and defining **diversity** through user actions were key topics. The group discussed democratizing information access and involving user preferences before deploying models.

Understanding user behaviors and creating accurate **user profiles** were emphasized. The discussion included improving existing user graphs (used to model and analyze connections between users, their activities, and the resources they interact with) and addressing privacy issues related to personalization.

Educating users on how to interact effectively with LLMs was deemed crucial. The group debated the benefits of long description queries and how to capture diverse user preferences to ensure the system aligns with a broad user base.

Personalization was recognized as having inherent risks, such as creating echo chambers. The group discussed whether the default mode should cater to general popular preferences or if users should be nudged with information from diverse contexts.

Addressing the **cold start problem** and the influence of search systems/LLMs on query writing were also key points. The extent to which ideal queries should be dictated by the system was debated.

Throughout the discussion, references to foundational works, such as Robert S. Taylor's study on question negotiation [Taylor, 1968] and information seeking in libraries, and Nicholas J. Belkin's concept of Anomalous States of Knowledge (ASK) [Belkin, 1980], provided a theoretical backdrop.

Establishing trust and creating mechanisms to escape the pitfalls of personalization were emphasized as critical components for the future development of LLMs. The group concluded that ethical practices, user education, and robust evaluation methods are essential for enhancing the effectiveness and reliability of LLMs.

5 Applications and Integration

Finally, we had a breakout group for discussing multifaceted applications and integration of LLMs. They began by comparing the merits of general-purpose LLMs with those fine-tuned for specific tasks, weighing the benefits of versatility against the precision of specialization.

The conversation naturally flowed into the realm of **multi-modal systems**, where information is conveyed through various formats such as text, images, and dynamic presentations. Participants debated the criteria for selecting these modalities, using examples like exploratory search, which might benefit from summaries, reference documents, and diverse outputs. They pondered whether LLMs should generate both text and images or focus solely on summarization, drawing parallels to Wikipedia's multi-modal approach.

The potential for LLMs to guide users along **learning paths** was another key topic. Designing interactions that support dynamic search was emphasized, contrasting with traditional recommendation systems for movies and music, which can sometimes lead users into uninteresting rabbit holes. Unlike these systems, LLMs require carefully designed feedback mechanisms to ensure relevance and engagement. Learning, they noted, is not just about acquiring information on a specific topic; broader context and serendipity play crucial roles. Multi-modality was seen as particularly beneficial in applications such as claim verification, where processing images alongside text can provide a more comprehensive understanding.

The group also discussed the limitations of chatbots as the primary interface for LLMs, suggesting that generating websites or other content might be more effective in certain contexts. They explored the concept of **mixed-initiative systems**, where the system takes some initiative by being proactive, and highlighted the challenges of controllability and the unpredictability of outputs when the system acts on behalf of the user.

Operational control and the availability of datasets for training these models were also discussed. Public datasets such as Microsoft's Common Objects in Context (COCO) [Lin et al., 2014] were mentioned, but the difficulty of experimenting and obtaining feedback was acknowledged. Risk assessment was highlighted as a critical first step in any LLM application, with accountability extending to all involved in the development process.

Ethical considerations, once again, were a significant part of the discussion. Examples such as the United States Transportation Security Administration's use of facial recognition and OpenAI's decision not to roll out emotion detection features in the European Union due to risk

illustrated the ethical dilemmas and potential stifling of development. The group debated the use of foundation models for tasks that currently require extensive experimentation and iteration.

The participants then discussed how **effective feedback mechanisms** are essential for refining LLMs. Ideally, models should immediately incorporate feedback, but current practices often involve RLHF or fine-tuning phases. The challenge of maintaining memory across chat sessions and deciding whether feedback should apply to the current session or persist indefinitely was also discussed.

The potential for extreme **personalization** in a privacy-preserving manner was seen as a significant benefit of LLM applications. Participants considered what context should be local versus cloud-based and noted the inconsistency in LLM behavior, which can make it difficult to restrict certain types of responses.

The group noted that there has been a shift in consumer expectations, with some tolerance for LLM errors. However, **reliability** remains an issue, as illustrated by the need for specific output formats in tasks such as the National Institute of Standards and Technology's Text REtrieval Conference (TREC) and the reluctance to answer certain types of questions.

The group debated whether the creators of LLMs should decide what is appropriate, referencing comprehensive experiments by organizations such as Anthropic. The concern was that a small number of people making content moderation decisions could impact everyone.

Overall, the discussion highlighted the complexities and challenges of integrating LLMs into various applications. Ethical considerations, robust feedback mechanisms, and careful design are essential to ensure effective and reliable user interactions, paving the way for the future development of LLMs.

6 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 19 participants (about 30% response rate), out of which 12 were students. Overall, the respondents found the workshop to be very useful (mean average usefulness rating of 4.21 out of 5). Following are some of the comments received from the participants:

- "Producing research questions around AI evaluation and Usability and seeing professionals in the field responding to them in the groups that I became member of during the workshop was a highlight that stood out for me."
- "I feel the breakout sessions were very informative and thought provoking. Also, the keynote was very relevant to my research domain. I found the event extremely insightful."
- "The interactive process of meeting people and engaging in conversations and learning diverse perspectives from faculty, other students and researchers from Microsoft research teams and other industries such as Amazon was also something I took home from the workshop."
- "It was a wonderful experience, useful experience. I just wish for more."
- "I loved meeting a lot of interesting people with complimentary interests, knowledge, and skills."

7 Next Steps

The success of the first workshop in this series organized last year inspired us to work on an edited book on this subject [White and Shah, 2025]. Thanks to the generous contributions of several scholars, we are on track to publish this book in early 2025 as part of the Springer IR Series. We believe this book, along with the lessons learned and shared through these two workshops will continue helping the IR community ask and answer important and interesting questions pertaining to task-based IR infused with GenAI.

Acknowledgments

This workshop was made possible by generous support from the US National Science Foundation (NSF) award IIS-2023924, Microsoft Research, and ACM SIGIR.

References

Nicholas J Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1):133–143, 1980.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the International Conference on Machine Learning*, pages 19274–19286, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

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

Ryen W White and Chirag Shah, editors. *Information Access in the Era of Generative AI*. Springer Nature, Switzerland, 2025.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.