Interactions with Generative Information Retrieval Systems

Mohammad Aliannejadi[1], Jacek Gwizdka[2], and Hamed Zamani[3]

[1] University of Amsterdam, The Netherlands
m.aliannejadi@uva.nl
[2] University of Texas at Austin, United States
jacekg@utexas.edu
[3] University of Massachusetts Amherst
zamani@cs.umass.edu

Draft of a chapter intended to appear in a forthcoming book on generative information retrieval, co-edited by Chirag Shah and Ryen White.

1 Introduction

At its core, information access and seeking is an interactive process. In existing search engines, interactions are limited to a few pre-defined actions, such as "re-query", "click on a document", "scrolling up/down", "going to the next result page", "leaving the search engine", etc. A major benefit of moving towards generative IR systems is enabling users with a richer expression of information need and feedback and free-form interactions in natural language and beyond. In other words, the actions users take are no longer limited by the clickable links and buttons available on the search engine result page and users can express themselves freely through natural language. This can go even beyond natural language, through images, videos, gestures, and sensors using multi-modal generative IR systems. This chapter briefly discusses the role of *interaction* in generative IR systems. We will first discuss different ways users can express their information needs by interacting with generative IR systems (Section 2). We then explain how users can provide explicit or implicit feedback to generative IR systems and how they can consume such feedback (Section 3). Next, we will cover how users interactively can refine retrieval results (Section 4). We will expand upon mixed-initiative interactions and discuss clarification and preference elicitation in more detail (Section 5). We then discuss proactive generative IR systems, including context-aware recommendation, following up past conversations, contributing to multi-party conversations, and feedback requests (Section 6). Providing explanation is another interaction type that we briefly discuss in this chapter (Section 7). We will also briefly describe multi-modal interactions in generative information retrieval (Section 8). Finally, we describe emerging frameworks and solutions for user interfaces with generative AI systems (Section 9). We conclude with a question: Will the myriad interaction possibilities afforded by generative AI systems be embraced by a broad user base, or will they remain merely a research curiosity?

2 Expressing Information Needs

An information need is what prompts users to seek information through various means, such as asking others, consulting printed resources, other media, or searching online. It arises from the awareness of a gap in a user's knowledge or understanding, necessitating the acquisition of information to bridge that gap [12, 25]. Bridging the gap helps to fulfill a specific purpose or goal, which is typically driven by a work task [13].

Prompt-based interactions with large language modelss (LLMs), and more broadly, multi-modal interactions with LLMs-based systems, provide an opportunity to fundamentally rethink the processes of searching for, finding, and using information, and how to support these activities. This fresh perspective has the potential to significantly transform user experience by enhancing how users express their information needs and achieve their goals.

We will frame our considerations using the information need model proposed by Robert Taylor in the 1960s [69, 70]. Taylor identified four levels of information need, each helping us to understand how users formulate questions in their minds, how they articulate them, and how they interact with information systems. The four levels of information need are: (1) **Visceral Need**: An inexpressible, unformulated need, felt as a vague sense of dissatisfaction, (2) **Conscious Need**: The user is aware of the need but cannot fully articulate it, (3) **Formalized Need**: The need can be clearly expressed and defined, (4) **Compromised Need**: The articulated need, as presented to an information system, often simplified or altered to fit the system's capabilities.

Traditional search systems typically support levels 3 and 4, but not 1 and 2. We believe that LLMs-based information access systems have the potential to support all four levels. Therefore, we use these four levels to structure our speculative list of ways users could be assisted in their interactions with generative AI. We will draw, in part, on well-known information seeking models [45, 40].

Support for **Visceral Need**: (1) *exploratory interactions*: provide users with broad, exploratory dialogue that might help users *clarify* their thoughts; suggest related topics to help users better understand and articulate their needs. This is an example of Clarification which we describe in Section 5. (2) *prompt suggestions*: offer prompt suggestions or follow-up questions to guide users towards more specific questions. This is an example of Proactive Interactions which we describe in Section 6.

Support for **Conscious Need**: (1) *partial expression of needs*: accept partially formed questions or statements of need; (2) *proactive support for refinement*: generate relevant information that helps users *refine* their understanding of what they're looking for; (3) *guided conversations*: engage in a dialogue to help users articulate their needs more precisely. We describe such approaches in more detail in Result Refinement (Section 4) and Proactive Interactions (Section 6).

Support for **Formalized Need**: (1) *direct queries*: respond directly to well-formulated questions with relevant information; (2) *structured responses*: provide detailed, structured responses that address specific aspects of the user's need;

(3) *advanced features*: offer options (e.g., filters) for further exploration or *clarification* based on the formalized need.

Finally, support for **Compromised Need**: (1) *flexibility of syntax*: offer flexibility to allow for iterative refinement of queries without strict syntax requirements; (2) *flexibility of language*: interpret and respond to a wide range of query formats, reducing the need for users to adapt their language significantly; (3) *feedback loop*: offer feedback on questions, suggesting modifications or alternative phrasings to better match the user's needs and system's capabilities. We describe such approaches in more detail in Proactive Feedback (Section 3).

Overall, the key advantages of LLMs in assisting users at all four levels of information need are:

- **Natural Language Processing**: LLMs can understand and respond to queries expressed in natural language, making them accessible even at the visceral and conscious need levels.
- **Contextual Understanding**: Advanced LLMs can maintain context over multiple interactions, allowing for a more nuanced exploration of information needs.
- **Broad Knowledge Base**: LLMs draw upon a vast range of information, potentially addressing needs across various domains and levels of specificity.
- **Adaptive Responses**: LLMs can tailor their responses based on the perceived level of the user's information need, understanding and responding to both simple and complex questions and providing more or less detail as appropriate.
- **Iterative Refinement**: The conversational nature of LLMs interactions allows users to refine their queries progressively, moving from visceral to formalized needs through dialogue
- **Enhanced Expressiveness**: Prompt-based interactions allow users to express their needs in more nuanced and detailed ways. Users can specify the format, tone, and depth of the information they seek, which can lead to more tailored and useful outputs. For instance, users can request summaries, detailed explanations, comparisons, or creative content, depending on their needs.

However, it's important to note that while LLMs offer powerful capabilities in addressing information needs across Taylor's levels, they also have limitations. They may sometimes provide plausible-sounding but incorrect information, lack true understanding of context beyond the immediate conversation, and cannot replace the critical thinking and expertise of human information professionals in complex scenarios.

While LLMs can offer enhanced capabilities for expressing information needs, they also introduce new challenges. Such as *Capability Gap*: users may struggle to formulate their intentions clearly and effectively, leading to a gap between what they want and what the LLMs provides. *Instruction Gap*: users need to learn how to craft effective prompts, which can involve understanding the LLMs's capabilities and limitations. *Evaluation of Outputs*: users must critically evaluate the LLMs's responses for accuracy and relevance, as LLMs can sometimes generate incorrect or misleading information. A recent paper introduced these three gaps and termed them collectively the *"Gulf of Envisioning"* [64].

In the following sections we address selected aspects of user-LLMs-based-system-interactions, 3 Proactive Feedback 4, Result Refinement 5 Clarification, 6 Proactive Interactions, 7 Explanation, and 8 Multi-Modal Interactions. In Section 9, User Interfaces, we discuss recent user interface frameworks and solutions.

3 Proactive Feedback

Recent developments in large language models have paved the path towards complex interactions between the user and the system. Generative IR models are able to satisfy user's information needs in multiple interaction turns. Among many possibilities, this enables users to provide feedback to the system. Feedback can be provided when is explicitly requested by the system, for example in the form of clarifying questions or preference elicitation [5, 82, 52, 49]. Section 5 discusses these aspects in more detail. Feedback can be also requested for assessing the quality of the system at the end or in the middle of a conversation. For instance, Amazon's Alexa Prize Challenge [51] has sought explicit rating feedback from users upon the completion of the conversation. Zamani et al. [85] introduce the possibility of improving this simple approach by asking context-aware questions for feedback and making natural language interactions within the conversation.

Feedback can be provided proactively by the user, which is the focus of this section. Perhaps the simplest type of feedback that users provide can be in the form of *repeating or reformulating the user's need in the same search session*. If detected, this often means that the user's need has not been addressed yet. Besides such simple scenarios, users may provide *explicit positive or negative feedback*. Explicit positive feedback are often easier to identify and interpret. They are often in the form of appreciation and hold a positive sentiment. Explicit negative feedback, on the other hand, is more challenging, more diverse, and perhaps more important for system designers as they help the system to improve and identify its limitations. Pointing out what parts of the system's response is inaccurate, why it is does not satisfy the user's needs, or expressing frustration and disappointment are examples of explicit negative feedback. Current state-of-the-art technologies often cannot successfully take advantage of explicit negative feedback and often limit themselves to acknowledging the system's limitations and apologizing from the users. There are huge potentials in successfully comprehending negative feedback from users.

In generative IR systems, grounding as relevance feedback is also relevant to the concept of explicit feedback. Trippas et al. [71] define grounding as discourse for the creation of mutual knowledge and beliefs. Examples include providing indirect feedback by reciting their interpretation of the results. This process can potentially enable CIS systems to better understand a user's awareness of the results, background knowledge, or information need.

We would like to highlight the potentials in providing implicit feedback as well. Progress in commercial (web) search engines is in debt to large-scale implicit feedback collected from user interactions, such as clicks, skipped results, dwell time, and cursor (mouse) movement. Implicit feedback in generative IR systems is more challenging, because it is more likely to deal with abandonment in each session. This means that users may leave the system as they receive the answer they want without providing any positive feedback. Alternatively, they may leave the system as they lose hope in getting the right answer from the system. Besides abandonment, changing topics and asking follow-up questions can be interpreted as an implicit feedback signal in generative IR. Interpreting these user behaviors is essential in improving generative IR systems.

Research in understanding and modeling implicit (negative) feedback is relatively sparse and future technologies can greatly benefit from further research in this space.

4 Result Refinement

4.1 An Overview of Result Refinement

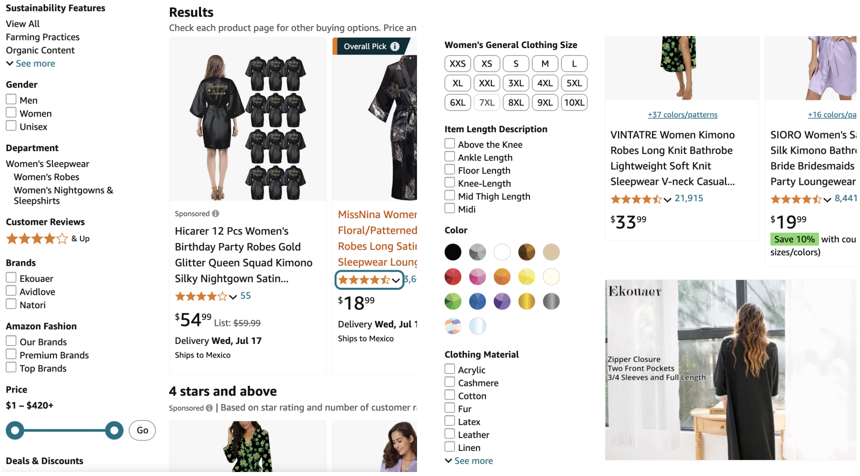
Result refinement is relatively understudied, compared to other modes of interaction in generative IR. Search result refinement has a long history of research in IR, especially in areas such as information filtering (e.g., recommender systems) where users access semi-structured information [16]. Figure 1 shows an example search result page from Amazon.com, where users are able to select certain attributes of the items (e.g., size) in the catalog to narrow down the results being presented to them. Search result refinement for semi-structured data is a relatively trivial task, as the refinement pane usually concerns the most important attributes of the items, given the query and the top item list. In the preference-based search literature, example-critiquing approaches have been explored [73], where the model suggests examples to the user, and with the user's feedback, it then models the user's preference. In conversational recommender systems, a similar approach is taken as part of the preference elicitation process [37]. In this process, the conversational system starts the conversation by asking the user's opinion about movies, aiming to optimize the decision space. A similar approach is taken in conversational product recommendation [97, 96]. In these works, the high-level idea is to extract important attributes from user reviews of products and model a probabilistic decision space. Then the conversational system takes a greedy approach in which, at every step, it aims to ask about an item attribute that minimizes the uncertainty of the decision space. Search result refinement is more challenging in web search, where the system deals with unstructured data. One of the earliest, simplest, and yet most effective ways is using vertical in the search result page [8]. Search result verticals divide the search results based on very high-level categories, such as images, videos, and news. Even though, very high level, it still can be considered as a naïve approach to refinement, as it approaches the user information from the result type. In most cases, the same user query can be satisfied with different modalities, which turns out to be one of the most important aspects of search, hence major commercial search engines still employ this approach. Finally, some early approaches tried to diversify, but also refine search results based on automatically extracted information facets. Faceted search [72] provides a means of navigation through topic facets for the users, enabling them to narrow down their information needs, as well as the search space. These early systems mainly relied on automatic facet extractors [36].

4.2 Technical Challenges

In the generative era, result refinement faces both algorithmic and interactive challenges.

Algorithmic challenges. As the items or documents are being represented using model parameters, refining the results based on a single attribute of the item is less trivial. To address this challenge, several works study controllable recommendation via disentanglement [17], where the goal is to represent items as separated attribute vectors instead of a single latent vector. Some of these attributes would be mapped to actual attributes in the catalog (e.g., color, style), or some latent attributes. LLMs have shown to be capable of extracting query facets, relying solely on their intrinsic knowledge [41]. However, as shown in the literature, LLMs are not yet capable of effectively grounding [63], which leads to suboptimal planning of LLMs utilizing their intrinsic knowledge to take the best next action. For example, for cases where humans would As the items or documents are being represented using model parameters, refining the results based on a single attribute of the item is less trivial. To address this challenge, several works study controllable recommendation via disentanglement [17], where the goal is to represent items as separated attribute vectors instead of a single latent vector. Some of these attributes would be mapped to actual attributes in the catalog (e.g., color, style), or some latent attributes. LLMs have shown to be capable of extracting query facets, relying solely on their intrinsic knowledge [41]. However, as shown in the literature, LLMs are not yet capable of effectively grounding [63], which leads to suboptimal planning of LLMs utilizing their intrinsic knowledge to take the best next action. For example, in conversations where most humans would ask for refinement, LLMs fail to take the same action.

Interactive challenges. As mentioned above, there has been research on various modes of refinement, i.e., search verticals, item attributes, faceted search, and example critique. While each of these modes has been utilized for a specific interaction medium (e.g., web search vs. conversational search), generative systems could potentially mix them. For example, prompting the user about their preferred search result modality, rather than making an assumption. Moreover, Chen et al. [14] review the interactive challenges of LLMs in the light of personalization, highlighting the importance of user–system interactions in result presentation, specifically refinement. Among other challenges, they refer to laborious data collection for training LLMs to be effective interactive systems, which can hinder the learning process.



(a) (b)

Fig. 1. Examples of search result refinement from Amazon.com. The refinement panes on the left help users browse through the search results.

5 Clarification

In a generative retrieval setting where the system aims to provide a comprehensive response to the user, whether in a conversational or web search setting, it is of utmost importance to ensure that the user's intent is predicted with high confidence. This is particularly critical, as in traditional web search scenarios, the system would diversify the list of results to ensure that various facets or interpretations of the query are covered in the top results [57]. However, in a generative scenario, usually, a single answer is provided to the user, limiting the information that can be exchanged between the user and the system.

5.1 An Overview of Search Clarification

Clarifying questions have been studied extensively [34] in the context of conversational question-answering [52], information-seeking conversations [5], and web search [82].

Another line of research studies the role of mixed-initiative interactions for user preference elicitation [49, 37]. The goal here is to understand the user preference when multiple documents (items) can be deemed relevant to their information need. Radlinski et al. [49] study this problem for movie recommendation, where the user information need is typically generic (e.g., "romantic movies") with multiple potentially relevant items. The dialogue system's goal in this setting is to engage in a conversation to elicit user preference in a more fine-grained way.

There has been a body of research studying the effect of mixed-initiative interventions such as clarifying questions on user experience [35, 84, 95, 98]. Kiesel et al. [35] study the effect of voice query clarification on user experience and find even in cases where the system performance is not improved, users have better experience. In web search, Zamani et al. [84] study the effect of incorporating a clarification pane on the search result page, implemented in Bing.com. Analyzing the click logs, they find that the clarification pane improves user experience. More specifically, among the seven templates they use to generate the clarifying questions, they find clear preference towards certain question templates in terms of user engagement. Zou et al. [95] study the effect of the clarification pane in the same setting in a controlled experimental setup where they introduce three quality levels and measure user satisfaction and performance. They find that asking a low-quality question in a search session risks lower user engagement with questions of higher quality in the same session. This finding was confirmed in a follow-up work [98].

User engagement (i.e., click-through rate) can be considered as a user-oriented quality measure of clarifying questions. Sekulic et al. [58, 60] extract various SERP- and document-based features to predict user engagement while interacting with clarifying questions in a web-based interface [83]. Rahmani et al. [50] study the effect of various query- and question-based features to predict user satisfaction in the MIMICS dataset [83] where they find, among others, a positive sentiment in the clarifying question leads to higher user satisfaction. Sekulic et al. [61] instead predicts the usefulness of clarifying questions in the retrieval pipeline. Following an early study on the effect of different types of clarifying questions on retrieval performance [38], they train a classifier to predict the usefulness of a clarifying question and its answer in the retrieval pipeline and incorporate it in the retrieval pipeline if only it is predicted to be useful.

5.2 Technical Challenges

Planning. While the early works in this area focused mainly on ranking clarifying questions from a pre-collected question bank [3, 4], more recent studies aim towards leveraging the generation power of LLMs to generate clarifying questions [88]. However, generative systems based entirely on LLMs are not effective in proactive interactions, especially in generating clarifying questions when necessary [23, 63]. Initial experiments reveal the power of LLMs in understanding the context of a query or a search session [1] and generate potential questions based on the context when prompted [21]; however, they fail at planning when to ask and which question to ask [21, 63]. Shaikh et al. [63] conduct a study where they compare human–human conversations with system–human conversations and find that LLMs fail at effectively planning when to ask clarifying questions in a conversation, even though they can generate high-quality questions if they are explicitly prompted to do so. Deng et al. [22] propose a proactive chain-of-thought approach to enhance the planning capability of LLMs such as ChatGPT and show that it has a considerable effect on their interaction capabilities.

Evaluation. Evaluating generative systems comes with various challenges. On top of that, evaluating interactive generative systems involves even more challenges as the user response to a system output is required. A line of research looks at simulating and modeling the user–system interactions in a mixed-initiative setting [90, 11, 55, 2, 10, 48, 59]. User simulation can be beneficial to generative IR models in two ways: (i) they provide a means for evaluating generated content, and (ii) they can be used for training. Zhang and Balog [90] propose a user simulator for conversational recommendation to evaluate the system performance. This is followed by the work done by Sekulic et al. [59] and Owoicho et al. [48] in using GPT-based models to simulate users in a mixed-initiative information-seeking conversational system where the main goal of the simulator is to provide an answer to a generated clarifying question. They show that such simulators can lead to reliable evaluation of conversational systems.

There are various considerations to take into account in simulating and evaluating interactive generative systems:

- User effort: In interacting with the system, users bear different levels of cognitive load, which can lead to user fatigue as the number of interactions increases.
- User information gain: To model the true value of a clarifying question in a conversation, we need to model both the gain and effort a clarifying question brings to the conversation [2, 10].
- Information nuggets: Information gain can be modeled by breaking the user's information need into information nuggets and measuring how much asking a certain clarifying question would help us provide further information nuggets to the user.
- User model: As proposed by Balog [11], an effective user simulator should have various components, including a user mental model. Realistically, a single user simulator does not cover the needs and behavior of the wide range of users interacting with the system.

6 Proactive Interactions

Typically, users initiate the interaction with a generative retrieval system, for example by submitting a chit-chat utterance, asking a question, or submitting an action request. In mixed-initiative conversational systems, the agent is also able to initiate the conversation. This is also called a *proactive*, system-initiative, or agent-initiative conversation. Existing generative AI systems are relatively under-developed when it comes to proactive interactions [43]. A major reason is that initiating a conversation by the system is not only challenging, but can also be risky; frequent and non-relevant proactive interactions may annoy users and hurt user satisfaction and trust [85]. Therefore, whether and when to initiate a proactive interaction are the key decisions a proactive CIS system should make.

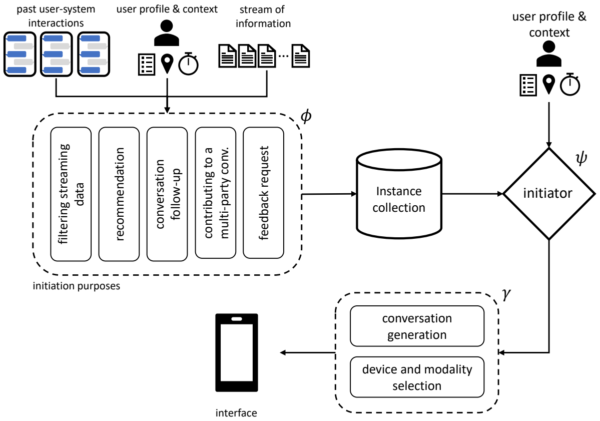


Fig. 2. A generic pipeline for conversation initiation in CIS systems by Wadhwa and Zamani [74].

6.1 An Overview of Proactive Generative Retrieval Systems

Wadhwa and Zamani [74] explored proactive conversational information access systems, discussing their challenges and opportunities. The authors introduced a taxonomy of proactive interactions, delineating three dimensions: (1) initiation moment (*when* to initiate a conversation), (2) initiation purpose (*why* to initiate a conversation), and (3) initiation means (*how* to initiative a conversation). They identified five purposes for initiating interactions: (1) filtering streaming information, (2) context-aware recommendation, (3) following up a past user-system conversation, (4) contributing to a multi-party human conversation, and (5) requesting feedback from users. A generic pipeline for these systems is depicted in Figure 2. In this pipeline, several algorithms constantly monitor the user's context and information streams to produce conversation initiation instances, which are stored in a database. A conversation initiator component then selects an appropriate instance based on the situation, initiating a fluent and accurate utterance. Figure 2 is sufficiently generic for illustrating proactive interactions in generative retrieval models and we use it to describe research and open questions in proactive retrieval in more detail.

Initiating a conversation through recommendation stands as one of the most common scenarios for proactive interaction. For instance, a conversational information access system might suggest an item based on the user's situational context, such as their location, time, and preferences. It is worth noting the distinction from traditional conversational recommendation setups, where users typically initiate the conversation by requesting specific items [66, 92]. Recent efforts in joint modeling of search and recommendation and developing unified information access systems [80, 81, 87] represent a step towards developing proactive, and thus mixed-initiative, systems in search and recommendation. However, proactive conversations extend beyond mere recommendations.

For example, Avula and Arguello [9] devised a system for conducting wizard-of-oz experiments, investigating proactive interactions during conversational collaborative search. This system could seamlessly integrate into collaborative platforms like Slack,[4] where during a collaborative search task, an external u ser (acting as a wizard) provides information. Though advancements in this area are nascent, there exists considerable potential for systems to initiate context-based conversations, engaging users and eliciting feedback.

Consider a scenario where a user employs a mapping application to navigate to a restaurant. Leveraging contextual cues, a proactive generative retrieval system could subsequently initiate a conversation upon the user's return journey, inquiring about their dining experience. Such interactions not only enhance user engagement but also facilitate feedback collection, aiding in profile refinement. Similarly, in situations where a user encounters difficulty in task completion, a conversational system could autonomously engage in conversation, offering assistance [85].

6.2 User Responses to Proactive Interactions

While in generative retrieval systems, users have the freedom to provide a natural language response in any form, they can be categorized as follows [74]:

- Null action: Users provide no response to the initiated conversation. It is important to note that null action should not necessarily be construed as negative feedback, as users may find the initiation useful but may not desire further engagement.
- Interruption or negation: Users respond in a manner consistent with terminating any further engagement by the generative retrieval system. It is perhaps safe to interpret such responses as negative feedback.
- Relevant response: Users provide a pertinent response to the initiated interaction, typically occurring when the interaction involves a question or solicits feedback.
- Postpone: Users respond to the initiated conversation and request the system to remind them at a later time.
- Critique or clarification-seeking response: Users engage further with the generative retrieval system, either seeking more information or critiquing existing engagement.
- Follow-up: Users provide a follow-up response to obtain additional information or perform actions related to the initiated conversation.
- Topic drift: Users respond but shift the topic of the initiated conversation.

[4] https://slack.com/

6.3 Technical Challenges

Here, we outline key technical hurdles in implementing the pipeline shown in Figure 2.

Producing System-Initiative Instances. The initial step in the system-initiation pipeline involves identifying reasons for initiating a conversation and generating a proactive instance. Proactive instances encapsulate all relevant information about a conversation, including its purpose, content, and context. This process entails addressing each initiation purpose component outlined in Figure 2. While some purposes, such as filtering streaming information and recommendation, have received attention in the literature, others like following up a past conversation or contributing to a multi-party conversation remain relatively unexplored. Thus, a major technical challenge lies in developing models capable of identifying the reasons for conversation initiation across various goals, including filtering information, recommendation, conversation follow-up, contributing to multi-party conversations, or requesting feedback.

Developing an Initiator Model. The subsequent step involves selecting a proactive instance from the instance collection using an initiator component. The primary challenge in this component stems from our limited understanding of the optimal moment to initiate a conversation. Consequently, future research should emphasize conducting user studies to explore the ideal timing for conversation initiation. Weak signals gleaned from user interactions with existing conversational systems, even those lacking proactive capabilities, could provide valuable insights. For instance, instances, where users initiate trivial conversations (e.g., out of boredom), could serve as noisy but potentially useful signals for predicting optimal conversation initiation moments. Machine learning models trained on situational context and user profiles could leverage such signals. Furthermore, interactive systems that log user interactions offer the opportunity to iteratively refine prediction accuracy based on user feedback.

Generating System-Initiative Utterances. The final step entails generating a (natural language) interaction based on a proactive instance and presenting it to the user. Techniques from dialogue systems and text generation research can be leveraged for this purpose. Since users typically do not anticipate proactive utterances, a notable technical challenge lies in providing context within the generated utterance to ensure user comprehension. This context could reference previous interactions with the system, user experiences, or explanations regarding the rationale behind initiating the conversation. Given that each instance is a structured data object, neural models designed for unstructured text generation from structured data, such as tables, could be potentially useful.

6.4 Evaluation of Proactive Systems

Assessing proactive generative IR systems poses significant challenges. While IR research has traditionally focused on creating collections for specific information-seeking tasks, these collections are typically based on predefined needs (e.g., TREC[5] tracks) or observations (e.g., clickthrough data). However, these evaluation methods do not readily apply to scenarios involving proactive interactions. Although evaluating proactive generative IR systems remains largely unexplored in the literature, we can envision two classes of evaluation methodologies: (1) modular evaluation, and (2) end-to-end evaluation.

In modular evaluation, the quality of each component in Figure 2 is evaluated in isolation? For example, how accurate is the initiator component in identifying opportune moments for proactive interactions? This methodology simplifies evaluation in proactive systems, but does not provide a complete picture of the overall performance of the system from the user's perspective, and does not reflect real-world complexities.

In end-to-end evaluation, one can explore both offline and online evaluation strategies. For offline evaluation, each instance would encompass all necessary information for the system at a given timestamp, including past user-system interactions, user profiles, situational contexts, and streams of new information. The model's performance would then be assessed based on the generated proactive interactions, if applicable. Crafting a single evaluation metric capable of capturing all facets of conversation initiation evaluation presents a challenge, necessitating further investigation. Recently, Samarinas and Zamani [56] introduced a large-scale benchmark for proactive interactions to ongoing multi-party human conversations and proposed normalized proactive discounted cumulative gain (npDCG) for end-to-end evaluation of such systems. In a separate investigation, Sen et al. [62] suggested evaluating proactive recommendation within search sessions by aggregating a correlation measure over the session. This measure assesses the relationship between the expected outcome—comprising the list of documents retrieved with a true user query—and the predicted outcome, representing the list of documents recommended by a proactive search system.

In the realm of online evaluation, conventional A/B tests can serve as a valuable tool for assessing the system's efficacy. Additionally, interpreting user feedback—both positive and negative—can provide valuable insights into system performance.

7 Explanation

Explanation can be seen as a critical tool in search result presentation in generative systems, as users are interested in comprehensive justification and explanation of the presented results [28, 14]. Also, it can lead to more user trust in the results, potentially aiding the user to distinguish between a low-quality and a high-quality response.

7.1 An Overview of Explanation in IR

Zhao et al. [94] provide a survey on the explainability of LLMs where they provide a taxonomy of explanations, together with methods for explaining Transformer-

[5] https://trec.nist.gov/

based LLMs. Also, they discuss various methods for evaluating explanations for both local and global explanations. Krishna et al. [39] show that not only are explanations useful in user–system interactions, but they also improve the performance of LLMs. They study automatic rationale generation in a chain of thought (CoT) manner. Deng et al. [24] show that rephrasing the user input leads to a better understanding of the user request which in turn results in better performance of the LLM, which is complementary to CoT reasoning. In their tutorial, Anand et al. [6, 7] review Transformer-based explanation generation. Zhang et al. [89] addresses search explainability via the lens of query understanding, where the system's task is to predict the user intent considering their query as input. LiEGe [79] explains all the documents in the ranking jointly using a listwise explanation generator.

Evaluating explanations is challenging. For free-text generations, human evaluation is employed. In other cases, because of a lack of explanation, proxy explanations such as clicks, query descriptions, query aspect annotation, and topic annotation can be used. For feature-based models, explanations are evaluated based on the effectiveness of predicted features. As for counterfactual explanations, model-based evaluation is employed.

7.2 Modes of Explanation in Generative IR

The main mode of explanation used in generative models is free-form text, where the model would further elaborate why the provided answer is relevant to the user's input. The explanation often consists of two major parts: (i) a further description of user information need, and (ii) an explanation of the reasons why the generated response is relevant to the user's input. The system has a limited information bandwidth and cannot present the users with multiple intents of their query. Therefore, describing what the system "thinks" the user wants helps the user understand whether the system understands their intent or not [89]. This type of explanation aims to ensure the user that their information need is properly understood by the system and can lead to increased trust in the system. Also, in case of misunderstanding the user's information need, it provides the opportunity for the user to realize what is missing in their input. This can be seen as similar to scanning the SERP by the user, through which the user would have an idea if the system understands their information need correctly.

Another form of explanation is to provide citations. This has been studied more extensively in the NLP community where the generated text attribution [29]. The URL citations are supposed to provide evidence of the source of information from the web. However, there are concerns regarding the quality of the citations, as there is no clear way of controlling the LLM to ground its responses on the cited page [91]. Citing source documents, while being useful as a form of explanation, still does not provide a comprehensive idea of the relevance of the source. Comparing it to a typical SERP where the users are exposed to the URLs of the results, users already have a quality perception by scanning through the page title, summary, and URL. Even though the LLM-based search interfaces aim to mimic this experience, it is not yet clear which parts of the generated response are extracted from the cited document. Moreover, it is not clear how much the system depends on its intrinsic knowledge (i.e., model parameters) vs. the retrieved document. Therefore, more research in this area is required to understand how much different techniques and modes of explanation affect the users' perception of quality and trust. One potential alternative is to treat the system as an information-gathering tool [53], rather than an information system. In such cases, the responses would look like "After searching the web, I found numerous sources of information about your query. Two of more trustworthy sources mention that ...." With such a response, not only does the user learn about the search space of the given query, but also they learn about the most important information extracted from the topic documents.

8 Multi-Modal Interactions

Research has demonstrated the advantageous role of multimodal signals in both keyword-based and recommendation-driven searches, spanning from contextual item recommendations [33, 76] to visual and multimedia recommendations [46]. These signals also address challenges like cold-start issues [18, 47, 15] and aiding in explaining and visualizing recommendation outcomes [68]. A recent survey by Deldjoo et al. [19] offers insights into the role of multimedia content in recommendation systems, delineating how such content—comprising audio, visual, and textual elements—enriches real-world recommendation challenges.

A significant challenge in Multimedia Information Systems lies in fusing multiple modalities to derive meaningful representations. Recent advancements in multi-modal large language models employ joint representation techniques to establish a latent space where multiple modality information can be compared. However, aligning content data like text and images is relatively straightforward compared to aligning content with user preferences such as ratings or social media data.

Deldjoo et al. [20] explored multi-modal conversational information-seeking tasks from multiple perspectives. They investigated (1) *Why* using multi-modal interactions, (2) *Which* tasks to support in multi-modal conversational systems, (3) *When* to integrate multiple modalities in conversations, and (4) *How* to research multiple modalities and conversations to enable multi-modal conversational information seeking. Deldjoo et al. [20] highlight the importance of each of these perspectives through a real-world example:

> Imagine a person is cycling along the road on their way to work. She is planning her day, including tasks from presenting a budget, hosting a new client, picking up their children after school, and making dinner. The cyclist passes a flower on the sideroad, which caught her eye and wanted to know what this plant is. Since she is cycling on a busy road, she quickly stops, takes a photo, and keeps riding. Meanwhile, she asks her earbuds to tell her which plant that was by a spoken query such as "what was that plant and is it edible?"

The authors argue that generative IR systems with multi-modal interactions and multi-modal sensors can accomplish the user's need in this and even more complex scenarios. Dealing with multi-modal interactions is a multidisciplinary topic, spanning across research areas from information retrieval, recommender systems, multi-media, human-computer interactions, computer vision, and even psychological and cognitive sciences. The intersection of the research areas that enable people to search for information through multi-modal conversations has not received the attention it deserves and it might partially be due to the complexity of the topic in terms of both modeling and evaluation. Prior work are mostly limited to two modalities (image and text), e.g., [67, 78], and further development in multi-modal foundation models [42, 26] and multi-modal retrieval-augmented generation models [54], is expected to speed up progress in this area.

9 User Interfaces

While in Section 2 we focused on general interaction methods to assist users in expressing their information needs when interacting with generative AI, in this section we review recent work on interaction techniques and user interfaces for information access with LLMs. The design space is huge, and it is still under-researched and poorly understood. For example, out of approximately 750 preprints related to LLMs published on arXiv in the field of Information Retrieval between 2020-2024, only 22 mentioned "user interface" in their abstracts.

New human-LLM interaction frameworks are only starting to emerge. For example, recent work [27] reviewed 73 papers published in HCI conferences since 2021 to investigate the dynamics of human-LLM interaction. Authors identified four key phases in the interaction flow and developed a taxonomy of four primary interaction modes. The four phases are: *planning* - before an interaction, *facilitating* - during an interaction, *iterating* - refining an interaction, and *testing* - testing an interaction. The interaction modes include: *Standard Prompting, User Interface, Context-based and Agent Facilitator.* The *User Interface* mode is of most interest to us as it enhances user interactions with LLMs beyond the conversational interface by improving input, output, iteration, and reasoning processes. This mode contains five approaches, which could be used separately or in combination. (1) *Structured prompt* approaches assist users in creating multi-component prompts, which could range from zero-shot to few-shot, and support specification of constraints. Tools like PromptMaker [30] combine prefixes, settings, and examples in prompt creation. (2) *Varying output* approaches allow users to specify output formats. Early examples like GenLine and GenForm [31] facilitate generation of user specified mixed outputs, such as HTML, JavaScript, and CSS code. User's control over output format allows for high level of personalization and, potentially, enhances consumption of information. (3) *Iteration of interaction* approaches include features such as debugging, error labeling, regenerating, and self-repairing, enabling users to refine their original prompts and workflows. BotDesigner [86], for instance, helps users identify and label errors within conversations and offers a "retry" button to regenerate outputs. (4) *Testing of interaction* facilitates the testing of various prompt variations, useful for quick testing of complex solutions. Tools like VISAR [93] use visual programming to enable rapid prototyping and testing of writing organization. (5) *UI to support reasoning* incorporates direct manipulation in the Chain-of-Thought process, allowing users to actively participate in and reorganize reasoning sequences. Other approaches in this area offer visual programming techniques, such as chain designs and mind maps and enable a more interactive and user-defined reasoning framework [32, 65, 93]. For example, Graphologue [32] introduced: (1) graphical diagrams which convert text-based responses from LLMs into diagrams; (2) graphical dialogues which enable graphical, non-linear dialogues between humans and LLMs; and (3) interactive diagrams which allow users to adjust graphical presentation, its complexity and submit context-specific prompts.

MacNeil et al. [44] explores three methods for integrating LLMs into user interfaces through a framework called Prompt Middleware. The three methods are: (1) *Static Prompts* are predefined prompts generated by experts through prompt engineering. They can be invoked by using UI elements (e.g., buttons), allowing users to send high-quality prompts to with minimal effort. This method leverages best practices but limits user control over prompt generation. (2) *Template-Based Prompts* involve generating prompts by filling in a template with options selected from the UI. The template integrates expertise and best practices, giving users more control through UI options. This method is exemplified by the FeedbackBuffet prototype, a writing assistant that uses template-based prompts to generate feedback on writing samples [44]. (3) *Free-Form Prompts*: This method grants users full control over the prompting process. Although challenging, it is beneficial in scenarios where complete control is desired.

Wang et al. [75] present a proactive interface design that addresses challenges users face in initializing and refining prompts, providing feedback to the system, and managing cognitive load. They describe three interaction techniques (*Perception Articulation, Prompt Suggestions, Conversation Explanation*) and how they can be supported by user interface elements. Perception articulation is supported by a pre-task questionnaire and main prompt template - the first supports information need at the visceral level, while the latter at the formalized level. Prompt suggestions are provided through supportive function tabs, which support conscious need. Conversation explanations are also delivered through supportive function tabs, with a feedback mechanism allowing users to rate the usefulness of these explanations. This feature supports compromised needs. Evaluation with participants demonstrated the effectiveness of these supportive functions in reducing cognitive load, guiding prompt refinement, and increasing user engagement. In interviews, participants appreciated the perception articulation functions for setting expectations and the conversation explanations for balancing expectations and satisfaction.

On one hand, the design space of user interfaces for LLMs offers a myriad of new interaction possibilities. On the other, taking advantage of the new possibilities can lead to complexity, which can make interfaces harder to comprehend and can overwhelm users. From the history of search interface evolution, we know that more complex search interfaces have not been widely accepted. For example, faceted search UIs led to a sharp learning curve and increased cognitive load [77]. History likes to repeat itself. Will it be the case with user interfaces for LLMs? Will the more complex interfaces for LLMs become only niche products?

10 Conclusions

As mentioned multiple times throughout this chapter, handling complex interaction types and modalities has been relatively under-explored and the authors find it a rich area of investment for the further development of generative information retrieval systems. This chapter pointed out prior work on various interaction types, from expressing information need to result refinement and mixed-initiative interactions, including clarification, feedback, and proactive interactions. Recent developments in (multi-modal) foundation models, including large language models, have paved the path towards better understanding complex user interactions, but we are still far from ideal generative information retrieval systems that can satisfy user needs efficiently, effectively, fairly, and robustly.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Office of Naval Research contract number N000142212688, and in part by NSF grant number 2143434. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

Bibliography

[1] Abbasiantaeb, Z., Yuan, Y., Kanoulas, E., Aliannejadi, M.: Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In: WSDM. pp. 8–17. ACM (2024)

[2] Aliannejadi, M., Azzopardi, L., Zamani, H., Kanoulas, E., Thomas, P., Craswell, N.: Analysing mixed initiatives and search strategies during conversational search. In: CIKM. pp. 16–26. ACM (2021)

[3] Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). CoRR **abs/2009.11352** (2020)

[4] Aliannejadi, M., Kiseleva, J., Chuklin, A., Dalton, J., Burtsev, M.: Building and evaluating open-domain dialogue corpora with clarifying questions. In: EMNLP (1). pp. 4473–4484. Association for Computational Linguistics (2021)

[5] Aliannejadi, M., Zamani, H., Crestani, F., Croft, W.B.: Asking clarifying questions in open-domain information-seeking conversations. In: SIGIR. pp. 475–484. ACM (2019)

[6] Anand, A., Lyu, L., Idahl, M., Wang, Y., Wallat, J., Zhang, Z.: Explainable information retrieval: A survey. CoRR **abs/2211.02405** (2022)

[7] Anand, A., Sen, P., Saha, S., Verma, M., Mitra, M.: Explainable information retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 3448–3451. ACM (2023)

[8] Arguello, J., Diaz, F., Callan, J.: Learning to aggregate vertical results into web search results. In: CIKM. pp. 201–210. ACM (2011)

[9] Avula, S., Arguello, J.: Wizard of oz interface to study system initiative for conversational search. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. p. 447–451. CHIIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3343413.3377941, https://doi-org.ezp.lib.unimelb.edu.au/10.1145/3343413.3377941

[10] Azzopardi, L., Aliannejadi, M., Kanoulas, E.: Towards building economic models of conversational search. In: ECIR (2). Lecture Notes in Computer Science, vol. 13186, pp. 31–38. Springer (2022)

[11] Balog, K.: Conversational AI from an information retrieval perspective: Remaining challenges and a case for user simulation. In: DESIRES. CEUR Workshop Proceedings, vol. 2950, pp. 80–90. CEUR-WS.org (2021)

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

[13] Byström, K., Hansen, P.: Conceptual framework for tasks in information studies. Journal of the American Society for Information Science and Technology **56**(10), 1050–1061 (2005). https://doi.org/10.1002/asi.20197

[14] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., Pu, Y., Lei, Y., Chen, X., Wang, X., Lian, D., Chen, E.: When large language models meet personalization: Perspectives of challenges and opportunities. CoRR **abs/2307.16376** (2023)

[15] Cui, P., Wang, Z., Su, Z.: What videos are similar with you?: Learning a common attributed representation for video recommendation. In: Hua, K.A., Rui, Y., Steinmetz, R., Hanjalic, A., Natsev, A., Zhu, W. (eds.) Proceedings of the ACM International Conference on Multimedia, MM '14,. pp. 597–606. ACM (2014)

[16] Cui, Z., Yu, F., Wu, S., Liu, Q., Wang, L.: Disentangled item representation for recommender systems. ACM Trans. Intell. Syst. Technol. **12**(2), 20:1–20:20 (2021)

[17] Cui, Z., Yu, F., Wu, S., Liu, Q., Wang, L.: Disentangled item representation for recommender systems. ACM Trans. Intell. Syst. Technol. **12**(2), 20:1–20:20 (2021)

[18] Deldjoo, Y., Dacrema, M.F., Constantin, M.G., Eghbal-zadeh, H., Cereda, S., Schedl, M., Ionescu, B., Cremonesi, P.: Movie genome: alleviating new item cold start in movie recommendation. User Model. User Adapt. Interact. **29**(2), 291–343 (2019)

[19] Deldjoo, Y., Schedl, M., Cremonesi, P., Pasi, G.: Recommender systems leveraging multimedia content. ACM Comput. Surv. **53**(5), 106:1–106:38 (2020)

[20] Deldjoo, Y., Trippas, J.R., Zamani, H.: Towards multi-modal conversational information seeking. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1577–1587. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). `https://doi.org/10.1145/3404835.3462806`, `https://doi.org/10.1145/3404835.3462806`

[21] Deng, Y., Lei, W., Lam, W., Chua, T.: A survey on proactive dialogue systems: Problems, methods, and prospects. In: IJCAI. pp. 6583–6591. ijcai.org (2023)

[22] Deng, Y., Liao, L., Chen, L., Wang, H., Lei, W., Chua, T.: Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In: EMNLP (Findings). pp. 10602–10621. Association for Computational Linguistics (2023)

[23] Deng, Y., Zhang, A., Lin, Y., Chen, X., Wen, J., Chua, T.: Large language model powered agents in the web. In: WWW (Companion Volume). pp. 1242–1245. ACM (2024)

[24] Deng, Y., Zhang, W., Chen, Z., Gu, Q.: Rephrase and respond: Let large language models ask better questions for themselves. CoRR **abs/2311.04205** (2023)

[25] DERVIN, B., NILAN, M.: Information needs and uses. Annual review of information science and technology **21**, 3–33 (1986)

[26] Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., Wen, J.: Wenlan 2.0: Make AI imagine via a multimodal foundation model. Nat. Commun. **13**(1) (2022)

[27] Gao, J., Gebreegziabher, S.A., Choo, K.T.W., Li, T.J.J., Perrault, S.T., Malone, T.W.: A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2024). https://doi.org/10.1145/3613905.3650786

[28] Gao, J., Wang, X., Wang, Y., Xie, X.: Explainable recommendation through attentive multi-view learning. In: AAAI. pp. 3622–3629. AAAI Press (2019)

[29] Huang, J., Chang, K.C.: Citation: A key to building responsible and accountable large language models. CoRR **abs/2307.02185** (2023)

[30] Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., Cai, C.J.: PromptMaker: Prompt-based Prototyping with Large Language Models. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–8. CHI EA '22, ACM (2022). https://doi.org/10.1145/3491101.3503564

[31] Jiang, E., Toh, E., Molina, A., Donsbach, A., Cai, C.J., Terry, M.: GenLine and GenForm: Two Tools for Interacting with Generative Language Models in a Code Editor. In: Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology. pp. 145–147. ACM (2021). https://doi.org/10.1145/3474349.3480209

[32] Jiang, P., Rayan, J., Dow, S.P., Xia, H.: Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–20. ACM (2023). https://doi.org/10.1145/3586183.3606737

[33] Kaminskas, M., Ricci, F., Schedl, M.: Location-aware music recommendation using auto-tagging and hybrid matching. In: Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013. pp. 17–24. ACM (2013)

[34] Keyvan, K., Huang, J.X.: How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. ACM Comput. Surv. **55**(6), 129:1–129:40 (2023)

[35] Kiesel, J., Bahrami, A., Stein, B., Anand, A., Hagen, M.: Toward voice query clarification. In: SIGIR. pp. 1257–1260. ACM (2018)

[36] Kong, W., Allan, J.: Extracting query facets from search results. In: SIGIR. pp. 93–102. ACM (2013)

[37] Kostric, I., Balog, K., Radlinski, F.: Generating usage-related questions for preference elicitation in conversational recommender systems. Trans. Recomm. Syst. **2**(2), 12:1–12:24 (2024)

[38] Krasakis, A.M., Aliannejadi, M., Voskarides, N., Kanoulas, E.: Analysing the effect of clarifying questions on document ranking in conversational search. In: ICTIR. pp. 129–132. ACM (2020)

[39] Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., Lakkaraju, H.: Post hoc explanations of language models can improve language models. In: Advances in Neural Information Processing Systems 36 (2023)

[40] Kuhlthau, C.C.: Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science **42**(5), 361–371 (1991-06)

[41] Lee, J., Kim, J.: Enhanced facet generation with LLM editing. In: LREC/COLING. pp. 5856–5865. ELRA and ICCL (2024)

[42] Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. Foundations and Trends® in Computer Graphics and Vision **16**(1-2), 1–214 (2024). `https://doi.org/10.1561/0600000110`, `http://dx.doi.org/10.1561/0600000110`

[43] Liao, L., Yang, G.H., Shah, C.: Proactive conversational agents in the post-chatgpt world. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3452–3455. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). `https://doi.org/10.1145/3539618.3594250`, `https://doi.org/10.1145/3539618.3594250`

[44] MacNeil, S., Tran, A., Kim, J., Huang, Z., Bernstein, S., Mogil, D.: Prompt Middleware: Mapping Prompts for Large Language Models to UI Affordances (2023). `https://doi.org/10.48550/arXiv.2307.01142`

[45] Marchionini, G.: Information Seeking in Electronic Environments. Cambridge University Press (1995)

[46] McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015. pp. 43–52. ACM (2015)

[47] Oramas, S., Nieto, O., Sordo, M., Serra, X.: A deep multimodal approach for cold-start music recommendation. In: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, DLRS@RecSys 2017, Como, Italy, August 27, 2017. pp. 32–37. ACM (2017)

[48] Owoicho, P., Sekulic, I., Aliannejadi, M., Dalton, J., Crestani, F.: Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In: SIGIR. pp. 632–642. ACM (2023)

[49] Radlinski, F., Balog, K., Byrne, B., Krishnamoorthi, K.: Coached conversational preference elicitation: A case study in understanding movie preferences. In: SIGdial. pp. 353–360. Association for Computational Linguistics (2019)

[50] Rahmani, H.A., Wang, X., Aliannejadi, M., Naghiaei, M., Yilmaz, E.: Clarifying the path to user satisfaction: An investigation into clarification usefulness. In: EACL (Findings). pp. 1266–1277. Association for Computational Linguistics (2024)

[51] Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., Pettigrue, A.: Conversational AI: the science behind the alexa prize. arXiv preprint 1801.03604 (2018)

[52] Rao, S., III, H.D.: Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In: ACL (1). pp. 2737–2746. Association for Computational Linguistics (2018)

[53] Ren, P., Liu, Z., Song, X., Tian, H., Chen, Z., Ren, Z., de Rijke, M.: Wizard of search engine: Access to information through conversations with search engines. In: SIGIR. pp. 533–543. ACM (2021)

[54] Salemi, A., Altmayer Pizzorno, J., Zamani, H.: A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 110–120. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591629, https://doi.org/10.1145/3539618.3591629

[55] Salle, A., Malmasi, S., Rokhlenko, O., Agichtein, E.: Cosearcher: studying the effectiveness of conversational search refinement and clarification through user simulation. Inf. Retr. J. **25**(2), 209–238 (2022)

[56] Samarinas, C., Zamani, H.: Procis: A benchmark for proactive retrieval in conversations. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024), (to)

[57] Santos, R.L.T., MacDonald, C., Ounis, I.: Search result diversification. Found. Trends Inf. Retr. **9**(1), 1–90 (2015)

[58] Sekulic, I., Aliannejadi, M., Crestani, F.: User engagement prediction for clarification in search. In: ECIR (1). Lecture Notes in Computer Science, vol. 12656, pp. 619–633. Springer (2021)

[59] Sekulic, I., Aliannejadi, M., Crestani, F.: Evaluating mixed-initiative conversational search systems via user simulation. In: WSDM. pp. 888–896. ACM (2022)

[60] Sekulic, I., Aliannejadi, M., Crestani, F.: Exploiting document-based features for clarification in conversational search. In: ECIR (1). Lecture Notes in Computer Science, vol. 13185, pp. 413–427. Springer (2022)

[61] Sekulic, I., Lajewska, W., Balog, K., Crestani, F.: Estimating the usefulness of clarifying questions and answers for conversational search. In: ECIR (3). Lecture Notes in Computer Science, vol. 14610, pp. 384–392. Springer (2024)

[62] Sen, P., Ganguly, D., Jones, G.: Procrastination is the thief of time: Evaluating the effectiveness of proactive search systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 1157–1160. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3210114, https://doi.org/10.1145/3209978.3210114

[63] Shaikh, O., Gligoric, K., Khetan, A., Gerstgrasser, M., Yang, D., Jurafsky, D.: Grounding gaps in language model generations. In: Duh, K., Gomez, H., Bethard, S. (eds.) Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6279–6296. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.naacl-long.348

[64] Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., Seifert, C.: Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–19. ACM (2024). https://doi.org/10.1145/3613904.3642754

[65] Suh, S., Min, B., Palani, S., Xia, H.: Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–18. ACM (2023). https://doi.org/10.1145/3586183.3606756

[66] Sun, Y., Zhang, Y.: Conversational recommender system. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 235–244. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3209978.3210002, https://doi.org/10.1145/3209978.3210002

[67] Sundar, A., Heck, L.: Multimodal conversational AI: A survey of datasets and approaches. In: Liu, B., Papangelis, A., Ultes, S., Rastogi, A., Chen, Y.N., Spithourakis, G., Nouri, E., Shi, W. (eds.) Proceedings of the 4th Workshop on NLP for Conversational AI. pp. 131–147. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.nlp4convai-1.12, https://aclanthology.org/2022.nlp4convai-1.12

[68] Tangseng, P., Okatani, T.: Toward explainable fashion recommendation. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020. pp. 2142–2151. IEEE (2020)

[69] Taylor, R.S.: The process of asking questions. American Documentation **13**(4), 391–396 (1962). https://doi.org/10.1002/asi.5090130405

[70] Taylor, R.S.: Question-negotiation and information seeking in libraries. College & Research Libraries **29**(3), 178–194 (1968). https://doi.org/10.5860/crl_29_03_178

[71] Trippas, J.R., Spina, D., Thomas, P., Sanderson, M., Joho, H., Cavedon, L.: Towards a model for spoken conversational search. Inf. Process. Manage. **57**(2) (mar 2020). https://doi.org/10.1016/j.ipm.2019.102162, https://doi.org/10.1016/j.ipm.2019.102162

[72] Tunkelang, D.: Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2009)

[73] Viappiani, P., Faltings, B., Pu, P.: Preference-based search using example-critiquing with suggestions. J. Artif. Intell. Res. **27**, 465–503 (2006)

[74] Wadhwa, S., Zamani, H.: Towards system-initiative conversational information seeking. In: Proceedings of the Second International Conference on Design of Experimental Search and Information Retrieval Systems. pp. 102–116. DESIRES '21, CSUR (2021)

[75] Wang, B., Liu, J., Karimnazarov, J., Thompson, N.: Task Supportive and Personalized Human-Large Language Model Interaction: A User Study. In: Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval. pp. 370–375 (2024). https://doi.org/10.1145/3627508.3638344

[76] Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., Liu, H.: What your images reveal: Exploiting visual contents for point-of-interest recommendation. In: Proceedings of the 26th International Conference on World Wide Web, WWW 2017. pp. 391–400. ACM (2017)

[77] Wilson, M.: Evaluating the Cognitive Impact of Search User Interface Design Decisions. In: Proceedings of the 1st European Workshop on Human-Computer Interaction and Information Retrieval. pp. 27–30 (2011), http://ceur-ws.org/Vol-763/

[78] Wu, Y., Macdonald, C., Ounis, I.: Multimodal conversational fashion recommendation with positive and negative natural-language feedback. In: Proceedings of the 4th Conference on Conversational User Interfaces. CUI '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3543829.3543837, https://doi.org/10.1145/3543829.3543837

[79] Yu, P., Rahimi, R., Allan, J.: Towards explainable search results: A listwise explanation generator. In: SIGIR. pp. 669–680. ACM (2022)

[80] Zamani, H., Croft, W.B.: Joint modeling and optimization of search and recommendation. In: Proceedings of the First International Conference on Design of Experimental Search and Information Retrieval Systems. pp. 36–41. DESIRES '18, CSUR (2020)

[81] Zamani, H., Croft, W.B.: Learning a joint search and recommendation model from user-item interactions. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 717–725. WSDM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3336191.3371818, https://doi.org/10.1145/3336191.3371818

[82] Zamani, H., Dumais, S.T., Craswell, N., Bennett, P.N., Lueck, G.: Generating clarifying questions for information retrieval. In: WWW. pp. 418–428. ACM / IW3C2 (2020)

[83] Zamani, H., Lueck, G., Chen, E., Quispe, R., Luu, F., Craswell, N.: MIMICS: A large-scale data collection for search clarification. In: CIKM. pp. 3189–3196. ACM (2020)

[84] Zamani, H., Mitra, B., Chen, E., Lueck, G., Diaz, F., Bennett, P.N., Craswell, N., Dumais, S.T.: Analyzing and learning from user interactions for search clarification. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1181–1190. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401160, https://doi.org/10.1145/3397271.3401160

[85] Zamani, H., Trippas, J.R., Dalton, J., Radlinski, F.: Conversational information seeking. Foundations and Trends® in Information Retrieval **17**(3-4), 244–456 (2023). https://doi.org/10.1561/1500000081, http://dx.doi.org/10.1561/1500000081

[86] Zamfirescu-Pereira, J., Wong, R.Y., Hartmann, B., Yang, Q.: Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In: Proceedings of the 2023 CHI Conference on Human Factors

in Computing Systems. pp. 1–21. CHI '23, Association for Computing Machinery (2023). https://doi.org/10.1145/3544548.3581388

[87] Zeng, H., Kallumadi, S., Alibadi, Z., Nogueira, R., Zamani, H.: A personalized dense retrieval framework for unified information access. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 121–130. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591626, https://doi.org/10.1145/3539618.3591626

[88] Zhang, Q., Naradowsky, J., Miyao, Y.: Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In: ACL (Findings). pp. 6665–6694. Association for Computational Linguistics (2023)

[89] Zhang, R., Guo, J., Fan, Y., Lan, Y., Cheng, X.: Query understanding via intent description generation. In: CIKM. pp. 1823–1832. ACM (2020)

[90] Zhang, S., Balog, K.: Evaluating conversational recommender systems via user simulation. In: KDD. pp. 1512–1520. ACM (2020)

[91] Zhang, W., Aliannejadi, M., Yuan, Y., Pei, J., Huang, J.H., Kanoulas, E.: Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics (2024), https://arxiv.org/abs/2406.15264

[92] Zhang, Y., Chen, X., Ai, Q., Yang, L., Croft, W.B.: Towards conversational search and recommendation: System ask, user respond. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 177–186. CIKM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3269206.3271776, http://doi.acm.org/10.1145/3269206.3271776

[93] Zhang, Z., Gao, J., Dhaliwal, R.S., Li, T.J.J.: VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. pp. 1–30 (2023). https://doi.org/10.1145/3586183.3606800

[94] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for large language models: A survey. CoRR **abs/2309.01029** (2023)

[95] Zou, J., Aliannejadi, M., Kanoulas, E., Pera, M.S., Liu, Y.: Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. ACM Trans. Inf. Syst. **41**(1), 16:1–16:25 (2023)

[96] Zou, J., Chen, Y., Kanoulas, E.: Towards question-based recommender systems. In: SIGIR. pp. 881–890. ACM (2020)

[97] Zou, J., Kanoulas, E.: Learning to ask: Question-based sequential bayesian product search. In: CIKM. pp. 369–378. ACM (2019)

[98] Zou, J., Sun, A., Long, C., Aliannejadi, M., Kanoulas, E.: Asking clarifying questions: To benefit or to disturb users in web search? Inf. Process. Manag. **60**(2), 103176 (2023)