Engineering conversational search systems

A review of applications, architectures, and functional components

Engineering Conversational Search Systems: A Review of Applications, Architectures, and Functional Components

Phillip Schneider[1], **Wessel Poelman**[2], **Michael Rovatsos**[3], **and Florian Matthes**[1]
[1]Technical University of Munich, Department of Computer Science, Germany
[2]KU Leuven, Department of Computer Science, Belgium
[3]The University of Edinburgh, School of Informatics, United Kingdom
{phillip.schneider, matthes}@tum.de
wessel.poelman@kuleuven.be
michael.rovatsos@ed.ac.uk

Abstract

Conversational search systems enable information retrieval via natural language interactions, with the goal of maximizing users' information gain over multiple dialogue turns. The increasing prevalence of conversational interfaces adopting this search paradigm challenges traditional information retrieval approaches, stressing the importance of better understanding the engineering process of developing these systems. We undertook a systematic literature review to investigate the links between theoretical studies and technical implementations of conversational search systems. Our review identifies real-world application scenarios, system architectures, and functional components. We consolidate our results by presenting a layered architecture framework and explaining the core functions of conversational search systems. Furthermore, we reflect on our findings in light of the rapid progress in large language models, discussing their capabilities, limitations, and directions for future research.

1 Introduction

Accessing information has always been one of the primary functions of computer systems. Early systems relied on command-line interfaces with a specific syntax for data retrieval. As search systems evolved, database query languages enabled more complex queries but required technical knowledge. Then, free-text search engines allowed users to enter keywords in natural language, with information typically displayed as a result page listing relevant items (Höchstötter and Lewandowski, 2009). In recent years, the evolution of search systems has continued in the direction of human-like dialogues.

Conversational search has emerged as a novel search paradigm, marking a shift from traditional search engines to interactive dialogues with intelligent agents (Radlinski and Craswell, 2017; Zhang et al., 2018). Many people have grown accustomed to using conversational interfaces like chatbots and voice assistants (Klopfenstein et al., 2017). The widespread usage of dialogue systems has changed how humans expect to interact with computers (McTear et al., 2016). Although modern conversational agents have impressive skill sets, their information-seeking capabilities are relatively limited and often confined to answering simple questions. As a consequence, there is a growing research interest in developing conversational search interfaces that go beyond simple query-response interactions by supporting more complex mixed-initiative dialogues, which is further fueled by the surging popularity of large language models (LLMs) and their integration into many kinds of search applications.

Even though the topic of conversational search is relatively new, its fundamental concepts can be traced back to early works from the natural language processing (NLP) and information retrieval fields. So far, this emerging topic has been approached from different angles. While some researchers focus on theories and conceptual aspects (Azzopardi et al., 2018), others conduct dialogue analyses and build prototypes to ground abstract models in empirical studies (Vakulenko et al., 2021a). Yet, despite the ample literature about required properties, many proposed systems are too complex to implement. This apparent gap highlights the need for a more holistic inspection that connects theoretical requirements with realizable functional components.

We conducted a systematic literature review investigating different aspects of conversational search systems (CSSs) to address this research gap. The three main contributions are as follows:
(1) We identify the conceptual system properties and suitable application scenarios of CSSs.
(2) We consolidate architectures from the literature into a layered architecture framework and elaborate on the core functional components of CSSs.
(3) We discuss the manifold implications for aug-

menting CSSs with LLMs, highlighting their potential capabilities, limitations, and risks.

2 Related Work

In the related research literature on systems for conversational information-seeking, three categories are usually distinguished: search, recommendation, and question-answering (QA) (Zamani et al., 2023). As the name suggests, CSSs actively involve users in the search process. Through multi-turn dialogues, users enter queries, locate information, examine results, or refine their search goals. In contrast to search systems, recommender systems usually rely on data about user preferences and past interaction histories to help with decision-making by providing personalized recommendations. QA systems have been an active area of research for many decades. Given a text corpus or knowledge base and a dialogue history, conversational QA systems aim to find answers to natural language questions (Vakulenko et al., 2021b). It is worth noting that the boundaries between conversational search, recommender, and QA systems are blurred and overlap. Although surveys exist on the two latter system categories (Jannach et al., 2021; Zaib et al., 2022), our literature review is dedicated to search-oriented conversational interfaces.

Despite the growing body of research on conversational search, related work, such as surveys or systematic literature reviews, remains scarce. The few studies we found tend to have a narrow topic focus on certain application domains or challenges. For example, the survey from Adatrao et al. (2023) gives an overview of conversational search applications in biomedicine. In a different study, Keyvan and Huang (2022) address the challenge of dealing with ambiguous queries. Another literature study from Gerritse et al. (2020) investigates problematic biases in personalized content that conversational search agents can exhibit. Yet another work by Kiesel et al. (2021) is a comprehensive survey on meta-information in search-oriented conversations.

To the best of our knowledge, we are the first to provide a system-centric review across the development process, ranging from conceptualizing core functions to implementing architectural components. Unlike the mentioned studies, we do not look into specific challenges or domains within conversational search but take on a broad engineering perspective. We summarize valuable insights regarding the design and development of CSSs for several application use cases. Additionally, we address the recent interest surrounding LLMs and their potential implications for engineering CSSs.

3 Method

We conducted our systematic review based on the guidelines from Kitchenham et al. (2004). Our study aims to shed light on the complex engineering process behind CSSs from initial system requirements to technical implementations by focusing on three key aspects: (1) definitions and proposed application scenarios to conceptualize the functional requirements of CSSs, (2) architectural elements suggested in the literature to effectively support these required system properties, and (3) core functions of CSSs discussed in the academic literature along with their implementations.

To obtain relevant publications, we devised a search string for querying six academic databases, as presented in Table 2 of Appendix A. The publication period was restricted to the time window between 2012 and 2022, yielding 212 candidate papers that predated the emergence of primarily LLM-based dialogue systems like ChatGPT (OpenAI, 2022). Two researchers screened the papers for relevance, selecting a final set of 51 papers. Additionally, they performed forward and backward snowballing to include recent papers from 2023 and 2024, mainly focusing on LLMs for CSSs.

4 Results

4.1 Definitions and Application Scenarios

The concept of conversational search is not uniformly defined in the literature. We found three main categories of definitions. System-oriented definitions describe conversational search referring to architectural components (Sa and Yuan, 2020; Vakulenko et al., 2021a). Dialogue-oriented definitions emphasize the specifics of the dialogue interaction (Radlinski and Craswell, 2017; Kiesel et al., 2021). Task-oriented definitions state tasks the system must complete (Zhang et al., 2018; Trippas et al., 2020). Despite focusing on different aspects, the analyzed definitions point out similar qualities to distinguish CSSs from traditional search approaches. These qualities are often related to the theoretical framework of Radlinski and Craswell (2017), which provides a structure and set of characteristics for designing and evaluating CSSs. In summary, we identified four reoccurring system properties from the analyzed papers. Firstly,

mixed-initiative interaction lets both user and system collaboratively steer the dialogue. Secondly, *mutual understanding* involves the system revealing its capabilities and helping users express their needs. Thirdly, *context awareness and memory* refers to the system's ability to gather information from its surroundings and conversation history to adapt dynamically. Lastly, *continuous refinement* denotes improving retrieval performance through direct feedback or learning from past interactions.

Search Modality. These system properties open up a wide range of use cases, but the suitability of conversational search depends on the search modality and search task. CSSs can support text-based, speech-based, or hybrid interaction modalities. Aliannejadi et al. (2021) analyze various modality types and discuss their impact on the user's information gain during conversations. The authors mention examples like voice interfaces as speech-only options for service hotlines, text-based systems that can be integrated into messaging platforms or web search engines, and multimodal systems, such as virtual assistants or smart speakers with screens to display visual information. Contrary to text-based interfaces, spoken CSSs work without screens and are highly accessible because they do not require any technical expertise. Yet, conveying search results solely through speech output can overwhelm users (Deldjoo et al., 2021). Moreover, two studies conducted by Xing et al. (2022) and Sa and Yuan (2020) indicate that different modalities influence the search behavior concerning the frequency of query reformulation or how long search results are examined. Although the majority of CSSs in the literature are predominantly uni-modal and text-based, Liao et al. (2021) note a growing trend towards multimodal systems.

The modality and the nature of the search task determine the appropriateness of conversational interaction. A conventional data lookup with a graphical user interface may be more efficient in scenarios where the information need can be easily expressed. Concerning more ambiguous scenarios where the search goal is multi-faceted, and the data structure complex, a free-form conversation with iterative clarifications, reasoning steps, and feedback loops becomes applicable for conversational search (Radlinski and Craswell, 2017). In support of this, Ren et al. (2021b) and Schneider et al. (2023a) argue that dialogue-based search is particularly effective for exploratory search goals that involve progressively narrowing down information items from unfamiliar information spaces (White and Roth, 2009). Other tasks for which the usefulness of conversational search was highlighted are sequential QA, learning about a new topic, asking for personal recommendations, or making plans (Anand et al., 2020).

Application Scenarios. In our analysis of conversational search scenarios, we identified several real-world application domains that have been explored. While business and health were the most popular domains, we observed a significant diversification in the last years, including aerospace, gastronomy, law, news media, public services, or tourism (Liao et al., 2021). For example, several researchers have studied product search in e-commerce scenarios for eliciting user preferences across multiple dialogue turns (Bi et al., 2019; Xiao et al., 2021). A study from Bickmore et al. (2016) proposed a CSS to support people with low health and computer literacy to find information about clinical trials for which they may be eligible. In the domain of news media, Schneider et al. (2023b) demonstrate the integration of knowledge graphs with conversational interfaces to enhance exploratory search of newspaper articles. They present a knowledge-driven dialogue system and, through a large-scale user study with 54 participants, evaluate its effectiveness and derive design implications regarding functional improvements. Liu et al. (2021) compared conversational versus traditional search in a legal case retrieval scenario, showing that users achieve higher satisfaction and success in the conversational approach, especially when they lack sufficient domain knowledge. We find that the analyzed domain-specific systems often help overcome the absence of prior background knowledge, facilitating users in initiating the search process. Alternatively, these systems can provide assistance when the interface's modality is restricted and does not support conventional search methods.

4.2 Architecture Framework

Once the application scenario and desired system requirements are defined, the subsequent steps in the engineering process are to transform theoretical properties into technical implementations. This refers to functional components and their integration as part of the system architecture. We identified over 20 system architectures from the literature and consolidated reoccurring elements into the

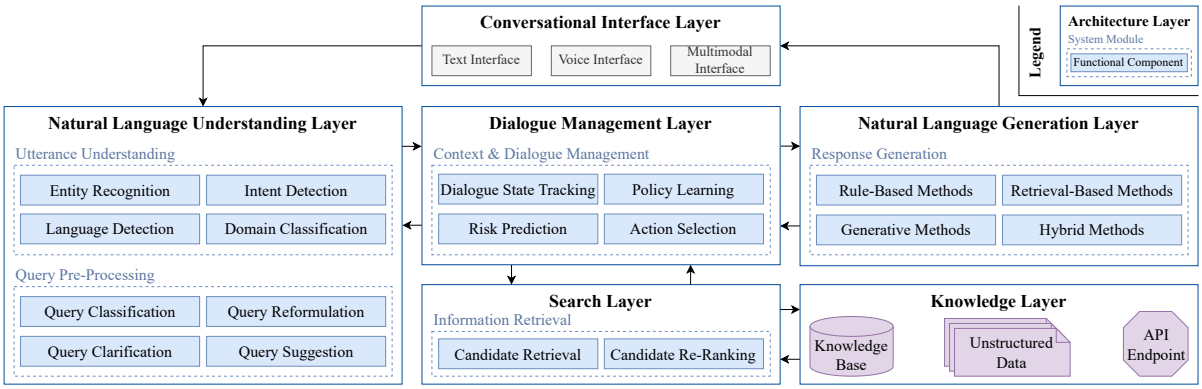

Figure 1: Architectural framework of conversational search systems.

generalized CSS architecture displayed in Figure 1. The proposed architecture adopts a layered architecture pattern, where each of the six layers performs a specific role within the CSS. The layers contain modules and functional components specifically designed for information-seeking purposes. For example, the *conversational interface layer* establishes the interaction channel between the system and the user. It receives user requests and presents search results depending on the modality. The three layers of *natural language understanding*, *dialogue management*, and *natural language generation* deal with processing input utterances, handling conversation logic, and producing responses as output. In CSSs, the correct understanding and meaningful pre-processing of user queries are essential to maximize the information gain. The *search layer*, in conjunction with the *knowledge layer*, performs search operations within the information space, ensuring access to various data structures. Possible data sources are corpora with unstructured text documents, application programming interfaces (APIs), or structured knowledge bases like knowledge graphs (Schneider et al., 2022). Data items can be stored in various databases, such as relational, graph, or vector databases, each with distinct benefits and drawbacks based on the data characteristics and application needs.

Modules group functional components and thus represent a specific functionality inside the layers. There is a separation of concerns among the modules, which deal only with logic pertinent to their respective layer. For instance, the query preprocessing module is a functionality from the language understanding layer, which enhances user queries through reformulation, clarification, suggestion, or other functions. The components perform specific tasks on the lowest abstraction level using NLP techniques. Implementing a component usually requires training NLP models that receive an input and classify, retrieve, or generate textual data, in some instances also structured data. Components can be implemented independently, requiring knowledge only of how they are connected to other components. While the displayed architecture encompasses all components encountered in the literature, implementations of a concrete CSS usually employ only a subset of these components. For example, reacting to user feedback is an essential function often mentioned in theoretical frameworks, but only a few studies implement it as part of an actual system (Bi et al., 2019; Wang and Ai, 2021). Since most architectures focus only on specific functional components like query suggestions or generating clarifying questions, there is a discrepancy between theoretical frameworks and practical implementations. Section 4.3 provides a more detailed overview of the various conversational search-specific core functions from the architectural components.

In line with common architectural patterns for dialogue systems, our proposed architecture follows a layered structure, separating functionality into different modules. We found that most analyzed implementations from the literature connect modules in a pipeline-based approach (Rojas Barahona et al., 2019; Mele et al., 2021; Alessio et al., 2023, *inter alia*). However, we observed a growing number of research works aiming to develop end-to-end approaches with transformer-based neural networks instead of classic NLP pipelines (Xiao et al., 2021; Ferreira et al., 2022). While end-to-end learning enables training a single model to represent target modules without the usual intermediate steps found in pipeline designs, these systems still depend on multiple task-specific modules and do

Functions	Example Studies	Datasets	Models	Access
Query classification	Aliannejadi et al. (2020)	TREC CAsT	BERT	♦
	Voskarides et al. (2020)	TREC CAsT, QuAC	BERT	✓
Query reformulation	Zhang et al. (2021)	TREC CAsT	HWE, T5	✓
	Yu et al. (2020)	TREC CAsT	GPT-2	✓
Query clarification	Zamani et al. (2020)	Bing search logs	BiLSTM	×
	Bi et al. (2021)	Qulac	BERT	♦
Query suggestion	Rosset et al. (2020)	Bing search logs	BERT, GPT-2	×
	Mustar et al. (2022)	TREC Session, MARCO, AOL logs	BERT, BART, T5	♦
Candidate retrieval	Xiong et al. (2020)	TREC DL, NQ, TriviaQA	ANCE	✓
	Lin et al. (2021)	TREC CAst, CANARD, MARCO	BERT	♦
Candidate re-ranking	Kumar and Callan (2020)	TREC CAsT	BERT	♦
	Mele et al. (2021)	TREC CAsT, ConvQ	BERT	✓
Knowledge-based response generation	Zhang et al. (2020)	WikiTableQuestions	T5, GPT-2	♦
	Ren et al. (2021a)	SaaC	PPG	✓

Table 1: Example studies, datasets, and implementations of the seven core functions in conversational search. Legend: ✓ = dataset(s) and system; ♦ = dataset only; × = not available.

not achieve a genuine end-to-end design, where only one model would handle all functionalities. To date, even the most advanced LLMs fail to integrate all functions without encountering issues, as we will discuss in more depth later on.

An example of a pipeline-based architecture is the open-source framework called *Macaw* from Zamani and Craswell (2020). It consists of three modules implemented in a generic form with replaceable NLP models. One module is responsible for query pre-processing with co-reference resolution and query reformulation or expansion, another for ranking documents with a retrieval model, and a third module for response generation. Two system proposals from Zhang et al. (2021) and Mele et al. (2021) have similar architectural components but additionally adopt a neural passage re-ranker for re-ordering results of the first-stage retrieval using a BERT model (Nogueira and Cho, 2019). Concerning end-to-end approaches, Xiao et al. (2021) introduce a CSS for online shopping, consisting of a sequence-to-sequence transformer for dialogue state tracking and a multi-head attention mechanism to match user queries to products. Comparable architectures from Ren et al. (2021a) and Ferreira et al. (2022) that aim to implement conversational search sub-tasks in an end-to-end manner also include transformers, such as BERT and T5, for passage re-ranking and response generation models.

Our presented architecture framework captures the fundamental aspects of CSSs in the research literature, and although there might be architectural adaptations to suit specific application scenarios with varying interface modalities and data structures, the body of six layers remains unchanged. The architecture offers flexibility in adding, removing, or replacing components within the modules.

4.3 Conversational Search Functions

This section elaborates on the seven core functions of CSSs mentioned in the architecture framework. Implementing these functions using NLP techniques is the most concrete step in the engineering process. Therefore, we review example studies that implement commonly used machine learning models (see Table 1) and list the most popular training and evaluation datasets in Table 3 of Appendix A. Despite being essential for conversational systems, some components like intent detection are not explicitly explained here as they are not specific to CSSs. While not all functions may be present in a given system or are combined, these main functions have been widely utilized and are treated as individual sub-tasks in the broader fields of conversational search and information retrieval. The order of paragraphs for each function roughly follows the processing steps needed to generate an output given an input turn in the conversation.

Query Classification. As part of the initial query pre-processing module, classifying the given query can benefit many subsequent system components. In conversational search scenarios, user requests may not be self-explanatory and ambiguous due to a lack of context. Researchers have approached this problem by classifying what type of question is being asked (Kia et al., 2020), determining the search domain of interest (Frummet et al., 2019; Hamzei et al., 2020), or deciding whether a (past) query is relevant in the context of the ongoing dialogue (Aliannejadi et al., 2020; Voskarides et al., 2020). Other system components can adapt according to classified queries, such as querying domain-

specific sources, discarding irrelevant utterances, or selecting relevant past utterances. The often-used TREC Conversational Assistant Track (CAsT) datasets contain many sessions where a user inquires about two subjects and later asks questions to compare the two. Classification can be used to select the previous relevant utterances.

Query Reformulation. Since a CSS is processing dialogue turns, it has to deal with many subtleties and challenges. Conversational search primarily deals with ambiguity and co-reference issues (Keyvan and Huang, 2022). Reformulating, also called rewriting, a query to an unambiguous and explicit form is often needed for effective information retrieval and to incorporate contextual information of an ongoing conversation. Numerous approaches incorporate transformer-based language models for this task (Ferreira et al., 2022). Either as a classifier to determine what terms have to be incorporated into the rewritten query (Mele et al., 2021), a sequence-to-sequence approach trained on *query – rewrite target* pairs (Zhang et al., 2021) or in a weakly-supervised fashion using LLMs (Yu et al., 2020). The following is a simple example of rewriting:

User:	Who is the director of Citizen Kane?
System:	Orson Welles is the director.
User:	Does he have children?
Rewrite:	Does ~~he~~ *Orson Welles* have children?

Query Clarification. When the system cannot resolve or interpret a query, it can take the initiative and ask the user for clarification. CSSs that can show initiative, such as proactively asking questions, are referred to as *mixed-initiative* systems. Different approaches for clarifying questions have been investigated, including template filling, sequence editing models, sequence-to-sequence models, and combinations of these methods. Template filling can be as straightforward as *"Did you mean X?"* for a misspelling or co-reference issue. Templates can cover many clarifying questions, but their specificity level is something to consider (Zamani et al., 2020). Sequence editing models are related to query rewriting; they choose a clarification question and rewrite it with information from the ongoing dialogue state (Zamani et al., 2023). Sequence-to-sequence approaches train models with *unclear query – clarifying question* pairs to predict fitting questions.

Asking a clarifying question is not always the best course of action. Systems have to ensure a user's patience or tolerance is not running out by asking too many questions (Bi et al., 2021). Controlling this 'risk' and the system's information need is a delicate balance. Current approaches implement functions that try to approximate the information gain and tolerance of a user (Salle et al., 2021; Wang and Ai, 2022). If the system wants to ask a clarifying question, it uses this function to decide whether it should proceed. This can be done for numerous reasons. Braslavski et al. (2017) provide a taxonomy of six clarification categories. Their categorical taxonomy is created from analyzing *community question-answering* websites but can be applied more generally.

Query Suggestion. CSSs can help users while they are still in their conversational turn by suggesting relevant queries or even (partial) answers while the interaction is ongoing (Aliannejadi et al., 2021; Keyvan and Huang, 2022). Search engines are a good example of this, where auto-complete is heavily used. Suggesting queries can possibly mitigate issues addressed by the previously mentioned system functions. If the system incorporates dialogue state information in the suggestions, it can provide unambiguous versions of an unclear query. Generating query suggestions is done in many ways, but all must deal with the query, dialogue state, and ranking-generated suggestions. An often-used approach is training a model to determine what to copy or generate from the dialogue state and input query to maximize the chance of a user picking the suggestion (Dehghani et al., 2017; Mustar et al., 2022). The generated queries can be ranked by the same or a separate model (Rosset et al., 2020).

Candidate Retrieval. Candidate retrieval fetches possibly relevant data items by producing a structured database query given the (pre-processed) user query or retrieving information from unstructured text collections. The latter approach falls into two general categories: sparse retrieval and dense retrieval (Gao et al., 2023). Sparse retrieval ranks documents with methods such as BM25 (Robertson and Zaragoza, 2009). These use sparse vectors encoding term occurrences in queries and documents, which can be used for retrieval directly, to perform pre-filtering of results (Vakulenko et al., 2021b; Zhang et al., 2021), or to represent model features (e.g., for re-ranking) (Cho et al., 2021). Although computationally efficient, the purely lexical approach of these methods limits them in dealing with synonyms, word order, and spelling mistakes.

Dense retrieval addresses these issues, which is often implemented as a *dual encoder* architecture, where one neural model encodes a document into a dense vector and another the (processed) query (Lin et al., 2021). These models are trained by jointly training these two encoders on labeled *query – relevant document* pairs. There are variations with additional encoding strategies, but the main idea stays the same (Ferreira et al., 2022).

Candidate Re-Ranking. Once the system has a set of possibly relevant candidate results for the current turn or utterance, the next step is to rank this set in order of informativeness. There are many approaches to re-ranking, with the most dominant one being some model that either classifies, scores, or re-orders a given input set (Ferreira et al., 2022). These models are either fine-tuned on explicitly labeled *query – relevant item* pairs (Zhang et al., 2021; Mele et al., 2021) or use some distance measure between (part of) the embedded query and (part of) the relevant document. These are the main building blocks of most implementations, but they can be combined into more elaborate setups. Kumar and Callan (2020), for instance, suggest *multi-view re-ranking*, where the system creates different embeddings of the input query. These *views* include information from dialogue history, relevant terms from the retrieved items, and the rewritten query, which get fused into the final ranking.

Knowledge-Based Response Generation. The final step of a turn in the conversational system is to present the response to the user in the form of natural language. As with information retrieval, natural language generation is a dedicated research field. As such, many distinct approaches and methods within CSSs exist. These are generally grouped according to three categories: the information type, generation method, and information source.

Information type refers to the response's structure based on the retrieved document(s) or information need. These include *short answer*, *long document retrieval*, *abstractive summarization* or *structured entities* (Zamani et al., 2023). For instance, a short factual question often does not require a large response (*"In what year did X happen?"*). In contrast, a query for an explanation might involve summarizing a relevant passage.

Different generation methods are used for these different answer types and can serve as a grouping of approaches. Some general methods include; template filling (Zhang et al., 2018), sequence-to-sequence methods (Ferreira et al., 2022) and weakly supervised approaches (Baheti et al., 2020). More elaborate approaches have a model choosing from where to copy a token in generating the response: a vocabulary, the input query, or the retrieved passage (Ren et al., 2021a,b).

Generation is also dependent on the information source being queried. Conversational search is generally done over a corpus of free text but can also be done over a knowledge graph (Kacupaj et al., 2022; Dutt et al., 2022) or other (semi-)structured information (Zhang et al., 2020). The source influences the choice of generation technique; verbalizing a sub-graph from a knowledge graph is considerably different from summarizing a text passage.

There are also hybrid methods that fuse information sources and generation methods. The most influential contribution in this area has been *retrieval-augmented generation* (Lewis et al., 2020; Shuster et al., 2021). These hybrid approaches try to balance the expressiveness and veracity of responses.

5 Discussion and Future Directions

The results from our review give insights into the engineering behind CSSs from abstract properties to realizable functional components. Against this background, our findings unveil a disruptive trend of adopting larger language models to integrate end-to-end functional components. Researchers have emphasized the benefits of streamlined NLP, reduced error propagation, and data-driven development. Hence, rather than reflecting on the numerous general challenges in the evaluation of CSSs, like Penha and Hauff (2020), we direct our focus toward discussing how LLMs can augment CSSs and the implications it has on their future evolution.

While most studies fine-tune language models (e.g., BERT or T5) on downstream tasks, there has been a recent surge of interest in using LLMs. By scaling up models to billions of parameters and training them on corpora with trillions of tokens, LLMs have demonstrated emergent capabilities and prowess in multi-task learning (Radford et al., 2019). A significant advantage of LLMs is prompt-based (or in-context) learning. Through carefully defined prompts, LLMs can perform multiple tasks without specific training or tuning (Liu et al., 2023). Furthermore, there has been a growing interest in optimizing LLMs for dialogue interactions by pre-training on conversations, instruction fine-tuning, and reinforcement learning from human feedback

(Thoppilan et al., 2022). The strengths of LLMs, such as their language understanding and ability to generate context-aware responses, make them highly complementary elements for CSSs.

Opportunities for Conversational Search. A rapidly growing body of new studies concentrates on advancing conversational search functions with LLMs. For instance, addressing the challenge of better understanding user queries, Anand et al. (2023) introduce a query formulation framework to replace multi-component pipelines with a single LLM. This model initially generates several machine intents for a user query, followed by options to accept, edit, or expand these intents until they align with the user's query intent. With a qualitative feasibility study, the authors show that the LLM-generated rewrites can improve the downstream retrieval performance. In related work, Mao et al. (2023) investigate different prompting and aggregation methods for performing few-shot conversational query reformulation with LLMs. They demonstrate that their approach outperforms state-of-the-art baselines by testing a GPT-3 model on CAsT'19 and '20 datasets. Another study from Chen et al. (2023) introduces a retrieval-based query rewriting approach, where an LLM leverages external knowledge from graphs with historical user-entity interactions and collaborative filtering. Ye et al. (2023) also demonstrate the potential of LLMs for query rewriting, showing that rewrites can significantly enhance retrieval performance in conversational search. Furthermore, LLMs can augment CSSs through semantic parsing and convert a natural language question into a structured database query. For example, Schneider et al. (2024a) evaluate how well different-sized LLMs perform in generating knowledge graph queries for conversational QA based on dialogues by comparing various prompting and fine-tuning techniques. Aside from query rewriting and semantic parsing, LLMs can also be effective for classifying query intents (Srinivasan et al., 2022) or generating clarification questions (Kuhn et al., 2023).

In addition to the natural language understanding layer, LLMs can augment the layers of dialogue management, search, and natural language generation. For example, Friedman et al. (2023) developed a system for conversational video search and recommendation powered by several LLMs based on the LaMDA model (Thoppilan et al., 2022). While one LLM is used as a dialogue management module, a second LLM acts as a re-ranker module. This LLM also generates explanations for its decisions. The authors discuss how a third LLM can be instructed to act as a user simulator for generating synthetic data for training and evaluation. Also focusing on synthetic data generation, a paper from Huang et al. (2023) introduces a framework called *CONVERSER* that uses LLMs to generate conversational queries given a passage in a retrieval corpus for training dense retrievers. This can significantly benefit conversational search by reducing the need for extensive and expensive data collection while maintaining high retrieval accuracy. Concerning knowledge-based text generation, LLMs have also proven to be effective for verbalizing semantic triples retrieved from graph-structured data, with performance improvements achievable through few-shot prompting, post-processing, and fine-tuning techniques (Schneider et al., 2024b). Another noteworthy approach from Sekulic et al. (2024) employed LLMs in conversational search for answer rewriting, proposing two strategies by either providing inline definitions of important entities or offering users the opportunity to learn more about entities. Human-based evaluations indicated a preference for the answers with inline definitions.

Challenges and Risks. Even though LLMs show great potential for conversational search, they have known shortcomings that must be considered. First, the sheer size of these models requires significant computational resources. Multiple graphical processing units are often necessary for enabling fast inference, a critical factor for conversational search applications that require responses in near real-time. The research community has been actively exploring solutions such as model distillation, model quantization, or low-rank adaptation to address these issues. Distillation involves compressing LLMs into smaller and more efficient versions (Shridhar et al., 2023). Model quantization is a technique where the floating point precision of model parameters is decreased, leading to smaller memory requirements and faster computations without significant performance loss (Xiao et al., 2023). Low-rank adaptation fine-tunes only a subset of the model's parameters rather than updating the entire parameter space (Hu et al., 2022).

Other major issues related to LLMs are hallucinating or omitting important information and a lack of transparency regarding the source from which the output was generated (Dou et al., 2022; Ji et al.,

2023; Xu et al., 2024). To mitigate these risks, scholars have looked into approaches to ground the generated outputs in trustworthy data sources and mechanisms to curate generated output. For example, Peng et al. (2023) introduce a framework for augmenting LLMs by first incorporating retrieved evidence from external knowledge as input context and then using LLM-generated feedback as instructions to revise responses. Through validation with two information-seeking tasks, the authors show that their approach reduces hallucinations while preserving fluency and usefulness. Another knowledge-enhancement method from Yang et al. (2023) fine-tunes a smaller LLM (Llama-7B) to learn domain-specific knowledge. This model is consulted to generate expert opinions that are used to enrich the prompt context of a bigger, general LLM (GPT-4) to improve its domain-specific QA capabilities. For a comprehensive survey of over 30 hallucination mitigation techniques, readers are referred to Tonmoy et al. (2024). Regardless, it must be noted that LLMs are nondeterministic by nature, making it challenging to ensure consistent and persistent knowledge during searches due to the inherent randomness in their text generation methods (Krishna et al., 2022; Mitchell et al., 2023).

Finally, there are efforts to develop software tools that address the reliability and safety of generated LLM output by adding programmable guardrails as well as logical control patterns. Popular tools that aid the development of LLM-based CSSs include *NeMo* (NVIDIA, 2023), *Guidance* (Microsoft, 2022), and *LangChain* (Chase, 2022). Other tools like *DeepEval* (Ip, 2024) can evaluate model bias, which is crucial since LLMs in conversational search can increase selective exposure and opinion polarization by fostering confirmatory querying behaviors (Sharma et al., 2024). In summary, ongoing research shows the potential of LLMs to advance the engineering of dialogue-based search systems with various approaches to mitigate their reliability issues. However, it is unlikely that LLMs will replace CSSs as a single end-to-end monolith in the foreseeable future. Instead, they are more likely to augment the modular structure of the proposed architecture framework.

6 Conclusion

We conducted a comprehensive review of engineering CSSs, establishing connections between theoretical application scenarios and technical implementations. Based on our analysis of existing architectures, we introduced a layered architecture framework and explained its functional core components. While it is essential to acknowledge that the field of conversational search is rapidly evolving, and complete coverage is unattainable, our framework provides a generalized architecture based on previously validated systems. The framework does not claim to be exhaustive but rather serves as a foundational starting point for designing and developing CSSs. Lastly, we discussed recent work on the capabilities and challenges of augmenting CSSs with LLMs. We outline where they fit into our proposed framework, which core functions they have been used for, and highlight promising directions for future research.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) Software Campus grant 01IS17049.

References

Naga Sai Krishna Adatrao, Gowtham Reddy Gadireddy, and Jiho Noh. 2023. A survey on conversational search and applications in biomedicine. In *Proceedings of the 2023 ACM Southeast Conference*, ACMSE 2023, page 78–88, New York, NY, USA. Association for Computing Machinery.

Marco Alessio, Guglielmo Faggioli, and Nicola Ferro. 2023. Decaf: A modular and extensible conversational search framework. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3075–3085, New York, NY, USA. Association for Computing Machinery.

Mohammad Aliannejadi, Leif Azzopardi, Hamed Zamani, Evangelos Kanoulas, Paul Thomas, and Nick Craswell. 2021. Analysing Mixed Initiatives and Search Strategies during Conversational Search. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 16–26, New York, NY, USA. Association for Computing Machinery.

Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing Evolution of Multi-Turn Conversations for Effective Answer Retrieval. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 33–42, New York, NY, USA. Association for Computing Machinery.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking

conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational search–a report from dagstuhl seminar 19461. *Dagstuhl Reports*, 9(11):34–83.

Avishek Anand, Venktesh V, Abhijit Anand, and Vinay Setty. 2023. Query Understanding in the Age of Large Language Models. In *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, New York, NY, USA. Association for Computing Machinery.

Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeff Dalton. 2018. Conceptualizing Agent-Human Interactions during the Conversational Search Process. In *2nd International Workshop on Conversational Approaches to Information Retrieval (CAIR'18)*, Ann Arbor, MI, USA.

Ashutosh Baheti, Alan Ritter, and Kevin Small. 2020. Fluent response generation for conversational question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 191–207, Online. Association for Computational Linguistics.

Keping Bi, Qingyao Ai, and W. Bruce Croft. 2021. Asking Clarifying Questions Based on Negative Feedback in Conversational Search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, pages 157–166, New York, NY, USA. Association for Computing Machinery.

Keping Bi, Qingyao Ai, Yongfeng Zhang, and W. Bruce Croft. 2019. Conversational Product Search Based on Negative Feedback. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 359–368, New York, NY, USA. Association for Computing Machinery.

Timothy W. Bickmore, Dina Utami, Robin Matsuyama, and Michael K. Paasche-Orlow. 2016. Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment. *Journal of Medical Internet Research*, 18(1):e1.

Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly? Analyzing Clarification Questions in CQA. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 345–348, New York, NY, USA. Association for Computing Machinery.

Harrison Chase. 2022. Langchain: An app development framework using large language models. *LangChain GitHub Repository*.

Zheng Chen, Ziyan Jiang, Fan Yang, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Aram Galstyan. 2023. Graph meets LLM: A novel approach to collaborative filtering for robust conversational understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 811–819, Singapore. Association for Computational Linguistics.

Eunah Cho, Ziyan Jiang, Jie Hao, Zheng Chen, Saurabh Gupta, Xing Fan, and Chenlei Guo. 2021. Personalized Search-based Query Rewrite System for Conversational AI. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 179–188, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 729–738, New York, NY, USA. Association for Computing Machinery.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624v1*.

Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. 2017. Learning to Attend, Copy, and Generate for Session-Based Query Suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1747–1756, New York, NY, USA. Association for Computing Machinery.

Yashar Deldjoo, Johanne R. Trippas, and Hamed Zamani. 2021. Towards Multi-Modal Conversational Information Seeking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1577–1587, New York, NY, USA. Association for Computing Machinery.

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC Complex Answer Retrieval Overview. In *TREC*.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Ritam Dutt, Kasturi Bhattacharjee, Rashmi Gangadharaiah, Dan Roth, and Carolyn Rose. 2022. PerKGQA: Question Answering over Personalized Knowledge Graphs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 253–268, Seattle, United States. Association for Computational Linguistics.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can You Unpack That? Learning to Rewrite Questions-in-Context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5917–5923, Hong Kong, China. Association for Computational Linguistics.

Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2022. Open-domain conversational search assistants: The Transformer is all you need. *Information Retrieval*, 25(2):123–148.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961v2*.

Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. In *Natural Language for Artificial Intelligence*.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural Approaches to Conversational Information Retrieval*. Springer International Publishing, Cham.

Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, pages 133–136, New York, NY, USA. Association for Computing Machinery.

Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. 2020. Place Questions and Human-Generated Answers: A Data Analysis Approach. In *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography, pages 3–19, Cham. Springer International Publishing.

Nadine Höchstötter and Dirk Lewandowski. 2009. What users see – Structures in search engine results pages. *Information Sciences*, 179(12):1796–1812.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Chao-Wei Huang, Chen-Yu Hsu, Tsu-Yuan Hsu, Chen-An Li, and Yun-Nung Chen. 2023. CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–387, Prague, Czechia. Association for Computational Linguistics.

Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436v3*.

Jeffrey Ip. 2024. Deepeval: The open-source llm evaluation framework. *Confident AI*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Computing Surveys*, 54(5):105:1–105:36.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. Contrastive representation learning for conversational question answering over knowledge graphs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 925–934, New York, NY, USA. Association for Computing Machinery.

Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. *ACM Computing Surveys*, 55(6):129:1–129:40.

Omid Mohammadi Kia, Mahmood Neshati, and Mahsa Soudi Alamdari. 2020. Open-Domain question classification and completion in conversational information search. In *2020 11th International Conference on Information and Knowledge Technology (IKT)*, pages 98–101.

Johannes Kiesel, Lars Meyer, Martin Potthast, and Benno Stein. 2021. Meta-Information in Conversational Search. *ACM Transactions on Information Systems*, 39(4):50:1–50:44.

Barbara A. Kitchenham, Tore Dyba, and Magne Jorgensen. 2004. Evidence-Based Software Engineering. In *Proceedings of the 26th International Conference on Software Engineering*, ICSE '04, pages 273–281, USA. IEEE Computer Society.

Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, pages 555–565, New York, NY, USA. Association for Computing Machinery.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Clam: Selective clarification for ambiguous questions with generative language models. In *ICML 2023 Workshop on Deployment Challenges for Generative AI*.

Vaibhav Kumar and Jamie Callan. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3971–3980, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An Environment for Multimodal Conversational Search across Multiple Domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 675–684, New York, NY, USA. Association for Computing Machinery.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized Query Embeddings for Conversational Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bulou Liu, Yueyue Wu, Yiqun Liu, Fan Zhang, Yunqiu Shao, Chenliang Li, Min Zhang, and Shaoping Ma. 2021. Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1622–1626, New York, NY, USA. Association for Computing Machinery.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1211–1225, Singapore. Association for Computational Linguistics.

Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface*. Springer International Publishing, Cham.

Ida Mele, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, and Ophir Frieder. 2021. Adaptive utterance rewriting for conversational search. *Information Processing and Management: an International Journal*, 58(6).

Microsoft. 2022. Guidance: A language for controlling large language models. *Microsoft GitHub Repository*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Agnès Mustar, Sylvain Lamprier, and Benjamin Piwowarski. 2022. On the Study of Transformers for Query Suggestion. *ACM Transactions on Information Systems*, 40(1):1–27.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled

reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085v5*.

NVIDIA. 2023. Nemo guardrails: An open-source toolkit for easily adding programmable guardrails to large language model-based conversational systems. *NVIDIA GitHub Repository*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813v3*.

Gustavo Penha and Claudia Hauff. 2020. Challenges in the evaluation of conversational search systems. In *KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption, KDD-Converse 2020*, volume 2666 of *CEUR Workshop Proceedings*. CEUR-WS. Virtual Workshop; KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption, KDD-Converse 2020, KDD-Converse 2020 ; Conference date: 24-08-2020 Through 24-08-2020.

Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 989–992.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA. Association for Computing Machinery.

Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021a. Conversations with Search Engines: SERP-based Conversational Response Generation. *ACM Transactions on Information Systems*, 39(4):47:1–47:29.

Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2021b. Wizard of Search Engine: Access to Information Through Conversations with Search Engines. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 533–543, New York, NY, USA. Association for Computing Machinery.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Lina M. Rojas Barahona, Pascal Bellec, Benoit Besset, Martinho Dossantos, Johannes Heinecke, Munshi Asadullah, Olivier Leblouch, Jeanyves. Lancien, Geraldine Damnati, Emmanuel Mory, and Frederic Herledan. 2019. Spoken Conversational Search for General Knowledge. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 110–113, Stockholm, Sweden. Association for Computational Linguistics.

Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading Conversational Search by Suggesting Useful Questions. In *Proceedings of The Web Conference 2020*, pages 1160–1170, Taipei Taiwan. ACM.

N. Sa and X.-J. Yuan. 2020. Challenges in conversational search: improving the system capabilities and guiding the search process. In *Proceedings of the 24th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2020)*, page 37–42.

Alexandre Salle, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2021. Studying the Effectiveness of Conversational Search Refinement Through User Simulation. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 587–602, Cham. Springer International Publishing.

Phillip Schneider, Anum Afzal, Juraj Vladika, Daniel Braun, and Florian Matthes. 2023a. Investigating conversational search behavior for domain exploration. In *Advances in Information Retrieval*, pages 608–616, Cham. Springer Nature Switzerland.

Phillip Schneider, Manuel Klettner, Kristiina Jokinen, Elena Simperl, and Florian Matthes. 2024a. Evaluating large language models in semantic parsing for conversational question answering over knowledge graphs. In *International Conference on Agents and Artificial Intelligence*.

Phillip Schneider, Manuel Klettner, Elena Simperl, and Florian Matthes. 2024b. A comparative analysis of conversational large language models in knowledge-based text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–367, St. Julian's, Malta. Association for Computational Linguistics.

Phillip Schneider, Nils Rehtanz, Kristiina Jokinen, and Florian Matthes. 2023b. From data to dialogue: Leveraging the structure of knowledge graphs for conversational exploratory search. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 609–619, Hong Kong, China. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Ivan Sekulic, Krisztian Balog, and Fabio Crestani. 2024. Towards self-contained answers: Entity-based answer rewriting in conversational search. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, CHIIR '24, page 209–218, New York, NY, USA. Association for Computing Machinery.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Michael Bendersky. 2022. QUILL: Query intent with large language models using retrieval augmentation and multi-stage distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 492–501, Abu Dhabi, UAE. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv:2201.08239*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing and Management: an International Journal*, 57(2).

Svitlana Vakulenko, Evangelos Kanoulas, and Maarten De Rijke. 2021a. A Large-scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. *ACM Transactions on Information Systems*, 39(4):49:1–49:32.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021b. Question Rewriting for Conversational Question Answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 355–363, New York, NY, USA. Association for Computing Machinery.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query Resolution for Conversational Search with Limited Supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 921–930, New York, NY, USA. Association for Computing Machinery.

Zhenduo Wang and Qingyao Ai. 2021. Controlling the Risk of Conversational Search via Reinforcement Learning. In *Proceedings of the Web Conference 2021*, WWW '21, pages 1968–1977, New York, NY, USA. Association for Computing Machinery.

Zhenduo Wang and Qingyao Ai. 2022. Simulating and Modeling the Risk of Conversational Search. *ACM Transactions on Information Systems*, 40(4):85:1–85:33.

Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query—Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Springer International Publishing, Cham.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Liqiang Xiao, Jun Ma, Xin Luna Dong, Pascual Martínez-Gómez, Nasser Zalmout, Chenwei Zhang, Tong Zhao, Hao He, and Yaohui Jin. 2021. End-to-End Conversational Search for Online Shopping with Utterance Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3477–3486, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhaopeng Xing, Xiaojun Yuan, and Javed Mostafa. 2022. Age-related Difference in Conversational Search Behavior: Preliminary Findings. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 259–265, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *Preprint*, arXiv:2007.00808.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–312, Singapore. Association for Computational Linguistics.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-Shot Generative Conversational Query Rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1933–1936, New York, NY, USA. Association for Computing Machinery.

Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Hamed Zamani and Nick Craswell. 2020. Macaw: An Extensible Conversational Information Seeking Platform. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 2193–2196, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, pages 418–428, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *arXiv preprint arXiv:2201.08808v2*.

Edwin Zhang, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. Chatty Goose: A Python Framework for Conversational Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2521–2525, New York, NY, USA. Association for Computing Machinery.

Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020. Summarizing and Exploring Tabular Data in Conversational Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1537–1540, New York, NY, USA. Association for Computing Machinery.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 177–186, New York, NY, USA. Association for Computing Machinery.

A Appendix

The Appendix provides supplementary material for our study, including a list of the six queried academic databases along with the applied search string (Table 2), as well as an overview of commonly used datasets for CSSs (Table 3).

Search String		
"conversational search" OR *"information-seeking dialogue"* OR *"conversational information retrieval"* OR *"conversational information-seeking"* OR *"information-seeking conversation"*		
Database	**Number of Papers**	**Database Link**
ACL Anthology	48	https://aclanthology.org
ACM Digital Library	101	https://dl.acm.org
IEEE Xplore	5	https://ieeexplore.ieee.org/Xplore
ScienceDirect	3	https://www.sciencedirect.com
Scopus	46	https://www.scopus.com
Web of Science	9	https://www.webofscience.com/wos/

Table 2: Search string and number of retrieved candidate papers per database.

Dataset	**Size**	**Source**	**Lang.**
Amazon Reviews (Ni et al., 2019)	9M products	Amazon product catalog	en
CANARD (Elgohary et al., 2019)	40K questions	QuAC dataset	en
CodeSearchNet (Husain et al., 2019)	2M code queries	GitHub repositories	en
ConvQ (Christmann et al., 2019)	11K QA dialogues	Wikipedia	en
DuConv (Wu et al., 2019)	30K dialogues	MTime.com	zh
MRQA (Fisch et al., 2019)	550K QA pairs	18 existing QA datasets	en
MS MARCO (Nguyen et al., 2016)	1M QA pairs	Bing search engine	en
MSDialog (Qu et al., 2018)	2K QA dialogues	Microsoft Community forum	en
Natural Questions (Kwiatkowski et al., 2019)	320K QA pairs	Google search engine	en
QuAC (Choi et al., 2018)	14K QA dialogues	Wikipedia	en
Qulac (Aliannejadi et al., 2019)	10K QA pairs	TREC Web Track	en
SaaC (Ren et al., 2021a)	748 QA pairs	TREC CAR, MS MARCO, WaPo news	en
TREC CAR (Dietz et al., 2017)	30M passages	Wikipedia	en
TREC CAsT (Dalton et al., 2020)	38M passages	TREC CAR, MS MARCO	en
TriviaQA (Joshi et al., 2017)	650K QA pairs	Wikipedia, quiz and trivia websites	en
WikiTableQuestions (Pasupat and Liang, 2015)	22K QA pairs	Wikipedia	en

Table 3: Commonly used datasets in the literature on conversational search systems.