Conversational Exploratory Search of Scholarly Publications Using Knowledge Graphs

Phillip Schneider and **Florian Matthes**
Technical University of Munich, Department of Computer Science, Germany
{phillip.schneider, matthes}@tum.de

Abstract

Traditional search methods primarily depend on string matches, while semantic search targets concept-based matches by recognizing underlying intents and contextual meanings of search terms. Semantic search is particularly beneficial for discovering scholarly publications where differences in vocabulary between users' search terms and document content are common, often yielding irrelevant search results. Many scholarly search engines have adopted knowledge graphs to represent semantic relations between authors, publications, and research concepts. However, users may face challenges when navigating these graphical search interfaces due to the complexity and volume of data, which impedes their ability to discover publications effectively. To address this problem, we developed a conversational search system for exploring scholarly publications using a knowledge graph. We outline the methodical approach for designing and implementing the proposed system, detailing its architecture and functional components. To assess the system's effectiveness, we employed various performance metrics and conducted a human evaluation with 40 participants, demonstrating how the conversational interface compares against a graphical interface with traditional text search. The findings from our evaluation provide practical insights for advancing the design of conversational search systems.

1 Introduction

Digital publication platforms have greatly expanded the accessibility of scholarly articles, offering an extensive range of publications that can be efficiently discovered through integrated search engines. These digital platforms provide researchers with access to millions of scholarly documents, encompassing conference papers, journal articles, workshop proceedings, and book chapters. The number of scholarly documents is growing exponentially, with estimates suggesting that it doubles approximately every 15 years (Bornmann et al., 2021). As the body of literature grows, traditional keyword-based search methods are becoming less effective at filtering and ranking relevant documents. These lexical methods rely heavily on well-formulated queries and otherwise yield irrelevant results. Researchers are often hindered by the so-called *vocabulary mismatch problem*, which manifests as differences between search terms and the terminology in the documents (Furnas et al., 1987). This issue is especially pronounced in open-ended and exploratory search scenarios, where users navigate unfamiliar information spaces. In such scenarios, users' incomplete knowledge of certain topics prevents them from formulating queries to access the information they need (Schneider et al., 2023a).

Reacting to the challenges posed by the high volume of scientific output, digital publication platforms have begun to make their search functionalities more intelligent by employing semantic search methods using natural language processing (NLP). These methods enable search engines to understand the context and intent behind user queries. Moving beyond exact keyword matches to semantic matches on a conceptual level can help identify relevant articles, even when different terms are used, thereby aiding users in discovering publications from subfields they are unfamiliar with. Complementing this, knowledge graphs (KGs) have established themselves as a versatile data structure for representing semantic relationships between interconnected entities like institutions, authors, topics, research fields, and other concepts.

Two popular examples of platforms that have incorporated KGs are Microsoft Academic (Wang et al., 2020a) and Semantic Scholar (Kinney et al., 2023). Microsoft Academic created the Microsoft Academic Graph, which supports semantic search, contextual query understanding, and personal recommendations. Similarly, the Semantic Scholar platform operates on the Semantic Scholar Academic Graph with more than 200 million papers. While these platforms offer a range of graph-based features and visualizations, they introduce usability hurdles by rendering graphical search interfaces more complex. Graphical interfaces can become less effective for exploratory search because of the added layers of complexity, causing users to experience cognitive overload (Sweller, 1988). This might be exemplified by the decline and eventual termination of Microsoft Academic in 2021, whose intricate interface likely has contributed to deterring users (Orduña Malea et al., 2014).

To address the complexity of graphical semantic search interfaces, we propose developing a conversational interface for discovering scholarly publications via dialogue interactions, leveraging a KG data structure. The emerging paradigm of conversational search promises to satisfy information needs using intuitive information-providing conversations while avoiding information overload (Radlinski and Craswell, 2017). Through interactions with conversational agents, users can resolve ambiguities, refine their queries, narrow down the relevant search space, and extract novel insights. Our study aims to provide insights into how conversational search systems integrated with KGs can enhance the discovery of publications, thereby improving navigation and information retrieval in the scholarly research landscape. To demonstrate the effectiveness of our developed system, we utilize the open-source corpus of the ACL Anthology as our data foundation. The source code, models, datasets, and questionnaires are made available via a public GitHub repository.[1] Our three main contributions are as follows: (1) We propose an architecture for integrating a conversational exploratory search system with a scholarly KG. (2) We implement the system by assembling different task-specific language models. (3) We conduct both a model-centric performance assessment and a human evaluation of the developed system with 40 participants.

2 Related Work

Conversational search systems are defined as conversational interfaces that support acquiring information through multi-turn dialogues. These systems progressed significantly in recent years, largely driven by the rapid adoption of large language models (LLMs). A growing body of research focuses on augmenting conversational search systems with LLMs (Schneider et al., 2024c), including utterance understanding (Kuhn et al., 2023), dialogue management (Friedman et al., 2023), knowledge retrieval (Lewis et al., 2020), and response generation (Sekulic et al., 2024; Schneider et al., 2024b). While LLMs hold great potential for conversational search systems, they are not without shortcomings. LLMs can hallucinate or omit crucial information, and their outputs often lack transparency regarding the source of generated content (Ji et al., 2023). In addition, LLMs are usually non-deterministic, posing challenges in ensuring consistent and correct knowledge due to the randomness in their text generation processes.

To mitigate issues of factuality and reliability in conversational systems and LLMs, researchers have studied using KGs to ground outputs in verifiable data sources. Integrating KGs with dialogue systems has long been a focus in the literature. KGs can replace static domain knowledge with dynamic ontologies and have shown effectiveness in conversational question answering (QA) (Christmann et al., 2019; Schneider et al., 2024a). By navigating entity nodes and relationships, KGs enhance conversational context and information exploration. Numerous studies support the use of KGs for improving utterance understanding, response generation, and dialogue management (Chen et al., 2019, 2023). While KGs are increasingly being combined with conversational agents in fields such as healthcare, law, and business, there remains a significant gap in their application within the scholarly domain. Thus far, Meloni et al. (2023b) are the only ones to propose combining a conversational agent with the *Academia/Industry DynAmics* (AIDA) KG (Angioni et al., 2021). Their AIDA chatbot focuses on QA by executing database queries to count, list, compare, or describe scholarly entities (e.g., authors or conferences), thereby offering senior researchers an overview of the research landscape through bibliometric data (Meloni et al., 2023a).

In contrast, our proposed system supports the conversational discovery of research articles for users with vague goals in open-ended search scenarios, building on insights from our previous work (Schneider et al., 2023b). Therefore, unlike the AIDA chatbot, which primarily assists senior researchers, our proposed system is designed to support exploratory information search for non-expert users looking to discover relevant publications on a given topic without requiring in-depth knowledge.

[1] Repository: github.com/philotron/CS-Scholarly-KG

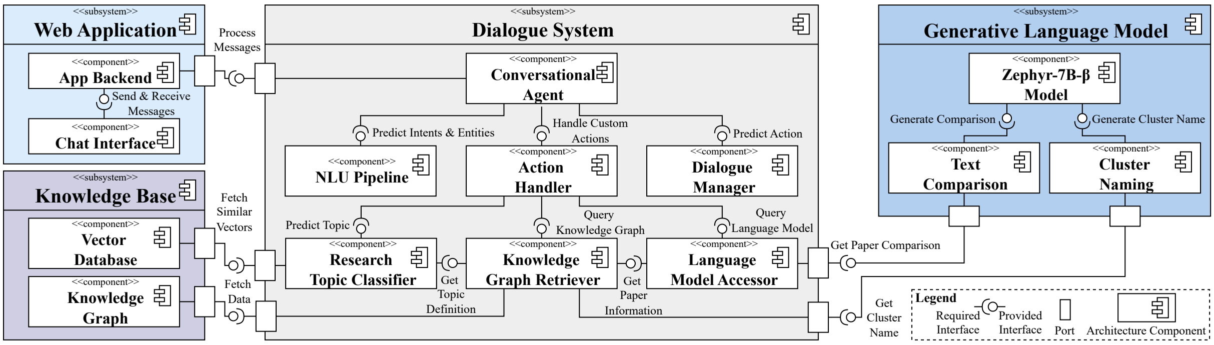


Figure 1: Architectural components of the conversational exploratory search system.

3 Conversational Search System

3.1 System Architecture

The developed dialogue system helps users narrow down relevant publications via a three-phase search process. An overview of the conversation flow is illustrated in Figure 4 in Appendix A. In the first phase, the system receives a short description of a search goal (e.g., "I want to study how people express their feelings on social media.") and then it assists users by recommending an appropriate NLP research topic to explore (e.g., *Emotion Analysis*). Second, users can iteratively choose thematic clusters of articles within the selected research topic (e.g., *Emotion Detection in Social Media Text*). Third, users are presented with a list of articles at the lowest cluster level, where they can compare papers based on summarized information from the abstracts. Finally, users have the option to either access links to the full texts of the papers or continue exploring other research topics, clusters, or articles. In Section 4.1, we will elaborate on the developed NLP models powering the three-phase search process through (1) topic classification, (2) text clustering, as well as (3) text summarization.

Figure 1 illustrates the system architecture, which is structured into four distinct subsystems. Each subsystem encompasses multiple components responsible for typical dialogue system functions. The front end is a conversational interface implemented as a web application subsystem using the open-source framework Streamlit.[2] It features a basic chat interface with a text input form and a scrollbar for the current dialogue history. User messages entered in the chat interface are sent to a dialogue system built with RASA, an open-source machine learning framework.[3] RASA supports the development of conversational agents with a natural language understanding (NLU) pipeline that recognizes intents and entities from user utterances. Based on these semantically parsed user utterances, the agent's dialogue manager, which takes into account dialogue states, dialogue policies, and conversation context, predicts the system's next actions. Aside from standard actions like producing a simple response, the agent connects with an action handler component to implement custom actions. One custom action is the KG retriever component. It enables the construction and execution of structured queries to retrieve data from the scholarly KG, such as abstracts, thematic clusters, or research topics. It connects with the knowledge base subsystem, which hosts the KG in a Neo4j property graph database.[4] For the query construction, extracted entities from user utterances are matched with those existing in the KG to fill out template queries.

In addition to the KG, the knowledge base subsystem hosts the open-source vector database Weaviate, which performs embedding-based similarity search.[5] Together with the KG, the vector database supports the research topic classifier component by finding the closest research topics from user inquiries. Another component that powers a custom action is the language model accessor, which provides a connecting endpoint to the generative language model subsystem. Inside this subsystem, we host the open-source LLM Zephyr-7B-Beta (Tunstall et al., 2023). The subsystem offers two inference endpoints for dynamic prompting. The endpoints are used to generate names for paper clusters and summarized paper comparisons.

The described system is deployed on three virtual machines (VMs) in a cloud environment. The first VM operates the dialogue system that interacts

[2] Streamlit: https://streamlit.io
[3] RASA: https://rasa.com
[4] Neo4j: https://neo4j.com
[5] Weaviate: https://weaviate.io

directly with users through the conversational interface. The second VM acts as a database server, while the third VM, equipped with a GPU (16 GB memory), hosts the large language model. Despite the architecture's various technical components depicted in Figure 1, the conversational interface hides the complexity of the underlying KG, providing a highly accessible search experience.

3.2 Knowledge Graph and Vector Database

To establish the data foundation for the conversational search system, we constructed a domain-specific KG with over 85,000 research articles sourced from the ACL Anthology.[6] A compact overview of the data schema with entity nodes and relations is presented in Figure 5 in Appendix A.

Sourcing articles from the ACL Anthology provided us with detailed metadata on authors, venues, and publication years that were automatically transformed into nodes and relations of the KG. In addition, we assigned each article to one or multiple research topics. To achieve this, we used a previously established taxonomy of NLP research topics from Schopf et al. (2023), which is organized as a two-level hierarchy with main topics and subtopics, along with topic names and human-written definitions. This taxonomy includes 12 main topics, such as *Text Generation* or *Sentiment Analysis*, and a total of 71 subtopics (e.g., *Question Generation* or *Emotion Analysis*). For classifying articles, we employed two fine-tuned language models for classifying publications: a SPECTER2-based model (Singh et al., 2023) for multi-label topic classification based on the used NLP taxonomy (Schopf et al., 2023; Schopf and Matthes, 2024), and a SciNCL-based model (Ostendorff et al., 2022) to classify if a publication is a survey paper consolidating information from several other publications. The taxonomy inside the KG is later applied to train a classification model that predicts a relevant NLP subtopic based on a described search goal. By providing topic definitions and listing related topics, the conversational agent can guide users through the NLP taxonomy.

While effective for broad classification, the two-level NLP taxonomy is not granular enough to account for thematic differences within a given subtopic. For example, the subtopic *Emotion Analysis* includes over 780 publications, which share only a few common characteristics. To have a more fine-grained filtering mechanism, we clustered papers based on their title and abstract content (e.g., specific techniques, application domains, or benchmark datasets). These thematic clusters are precomputed and modeled as nodes in the KG. The clustering and cluster naming methods will be discussed in more detail in Section 4.1.

In addition to the KG database, we installed a vector database that supports various embedding models and similarity metrics, making it ideal for efficiently ranking semantically similar documents. We employed the SPECTER2 embedding model (Singh et al., 2023) for generating vectors from papers' titles and abstracts and used them for the mentioned research topic classification, mapping NLP topics from the taxonomy to user requests during the search dialogue in real-time. A document identifier in the vector database links these embeddings to the papers in the KG. As a last construction step, we further enriched the KG with metadata from the Semantic Scholar API, including one-sentence *too long; didn't read* (TLDR) summaries, citation counts, and publication references.

4 Results and Discussion

4.1 Model Training and Evaluation

Research Topic Classification. In the following sections, we report the results of training and evaluating the NLP models that underpin the three-phase search process of our developed dialogue system: (1) research topic classification, (2) article text clustering, and (3) comparative article summarization. The first phase involves classifying an uttered search goal or problem description into a fitting NLP research topic. This is especially helpful for users in exploratory search settings because they may not be familiar with all existing fields of study and struggle to phrase their queries using the correct terminology. Due to the absence of datasets that map search goals expressed in layman's terms to NLP topics, we created a synthetic multi-class dataset using GPT-3.5-Turbo (version: 0613) (OpenAI, 2022). We prompted the LLM to generate questions on the 12 main topics in our taxonomy using three distinct personas: a computer science student with only peripheral NLP knowledge, a businessperson with practical experience of NLP tools but minimal technical expertise, and a non-technical, non-academic individual whose technology use is limited to basic tasks. Persona-specific prompting yielded diverse inquiries in lay-

[6] ACL Anthology: https://aclanthology.org

man's language. For example, the question "How are computers able to respond when we ask them questions?" was generated for the topic *Natural Language Interfaces*. To account for questions unrelated to NLP, we also generated a set of out-of-scope questions such as "Who discovered the laws of thermodynamics?" Following a quality inspection of the synthetically produced questions, we assembled a training dataset of 1601 examples, consisting of 120 questions for each of the 12 topics and 161 general questions. We also derived a test dataset containing 364 examples with a balanced class distribution similar to the training dataset.

In our experiments, we evaluated three classification approaches: vector similarity search, prompting a LLM (GPT-3.5-Turbo), and few-shot fine-tuning of a transformer model with the SetFit framework (Tunstall et al., 2022). Concerning the vector search approach with the SPECTER2 model, we measured the cosine similarity to compare vectors of embedded user queries with paper embeddings in our vector database to retrieve the 100 most similar papers. We found that a similarity threshold below 77% effectively filters out the non-NLP-related questions. Using the scholarly KG, we aggregated linked topics for these papers and predicted the most frequent topic as output class. For the LLM approach, we crafted a zero-shot prompt for GPT-3.5-Turbo, provided in Appendix A, which instructed the LLM to classify the appropriate topic from the list of 12 main topics or answer with "None" if the question was not related to NLP. Moreover, we tested the SetFit approach for fine-tuning the sentence transformer model multi-qa-MiniLM-L6-cos-v1 (Wang et al., 2020b). We trained for 3 epochs, a batch size of 16, and 30 SetFit iterations for contrastive learning.

Figure 2 illustrates the classification performance for each approach. While vector search achieved a macro F1-score below 0.50., GPT-3.5-Turbo achieved a score near 0.75; however, it exhibited a bias toward particular topics, leading to overprediction and incorrectly classifying general questions as NLP topics. The SetFit model demonstrated superior performance over the two other approaches with a score of 0.95. Consequently, we implemented the topic classifier component with this fine-tuned model for main topic classification in combination with similarity search for classifying the subtopic. This allows a more nuanced classification of user queries into subtopics, given the more detailed information in the paper abstracts.

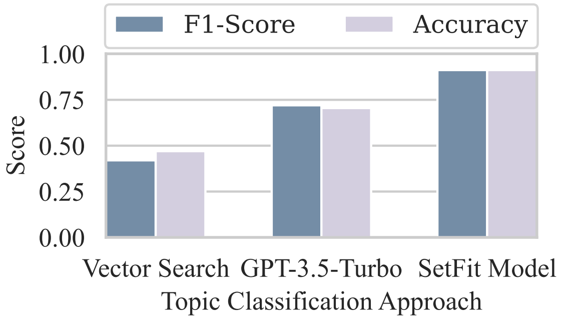


Figure 2: Comparison of accuracy and F1-scores for three topic classification approaches.

Article Text Clustering. After selecting an NLP subtopic, the conversational search agent guides users by presenting clusters of papers to further narrow down the search space. We tried out various clustering methods for thematically grouping similar papers because listing all papers within one subtopic at once is impractical. Selecting thematic clusters makes it easier for users to find relevant papers by iteratively choosing smaller, more specific clusters. Our experiments indicated that agglomerative clustering, a hierarchical bottom-up clustering approach, was the most effective in producing mutually exclusive clusters at each hierarchy level (Murtagh and Contreras, 2012). We employed the SPECTER2 embeddings of the publications, which were the same ones used for the similarity search as part of the research topic classification.

Initially, we used a distance threshold of 10, resulting in clusters averaging 15 publications each. However, this led to too many clusters inside a given research topic, making cluster selection very cumbersome. The distance threshold represents the maximum distance within which items are clustered together. To improve the clustering, we adopted an iterative hierarchical approach. We progressively decreased the distance threshold at each cluster level, keeping the number of clusters small while increasing the paper similarity within each subcluster to facilitate user navigation. Clustering stopped when fewer than 10 publications remained per cluster, ensuring a user-friendly number to display. Overall, we constructed a granular hierarchy of 47,035 thematic clusters, which were modeled as nodes in the constructed scholarly KG.

Next, it was necessary to assign human-readable cluster names to help users identify relevant clusters during the conversational search interaction. We applied a term frequency-inverse document frequency (TFIDF) vectorizer to extract important words from the titles of all publications within each cluster. Through several experiments, we found that an n-gram range of (2,5), considering sequences of two to five consecutive words, yielded good cluster names. However, we observed that a few clusters had identical names. To resolve this, we performed another cluster naming iteration, taking into account all previous names. If a name was repeating, we selected the second or third most relevant TFIDF label to ensure unique names. While the TFIDF-derived names were readable, they often contained too much detail and domain-specific words, rendering them less accessible to non-expert users. To make cluster names more understandable, the aforementioned Zephyr-7B-Beta LLM was applied. Within the LLM subsystem in our architecture, a specialized component was developed for cluster naming. A dynamic prompt, detailed in the GitHub repository, was created to transform an existing TFIDF cluster name alongside five randomly selected paper titles from a chosen cluster into a more comprehensible cluster name (e.g., *Emotion Detection in Social Media Text* or *Extraction of Concept Maps for Multi-Document Summarization*). To minimize response latency during a conversation, names for clusters and subclusters were pre-computed and stored in the scholarly KG.

Comparative Text Summarization. In the last phase of the conversational search process, users can compare papers listed at the lowest cluster level. Although these papers are already thematically related, the comparison allows users to discern specific similarities and differences, aiding in determining which paper to read more thoroughly. The language model subsystem allows for the summarization of objectives and results of two selected papers, which are generated in real-time upon request with Zephyr-7B-Beta. Given that injecting full abstracts can impede the LLM's ability to accurately detect objectives and results, only relevant portions of the abstract are provided in a dynamic prompt, which has been shown to reduce hallucinated outputs (Martino et al., 2023).

To this end, we first classified abstract sentences that discuss objectives or results using SciBERT, a language model pre-trained on scientific text (Beltagy et al., 2019). We fine-tuned SciBERT on a labeled dataset from Gonçalves et al. (2020), including 500 computer science abstracts and 3,287 sentences classified as background, methods, objectives, results, or conclusions. After hyperparameter optimization, our fine-tuned model achieved an F1-score of 75.39%, which is around one percentage point higher than the model from Gonçalves et al. (2020) with 74.60%. Finally, we applied our model to all the publication abstracts in our KG and stored the classified objectives and results sentences accordingly. More details about our fine-tuned model are available in the repository. A dynamic LLM prompt for text summarization was crafted, as shown in Table 5 in Appendix A. Two researchers manually assessed the generated comparisons. Initial experiments with other models, such as Falcon-7B and Llama-2-7B, showed that these models were less attuned to following the instruction, often producing hallucinated content or excessively verbose responses, making them unsuitable for conversational interactions. As a result, we selected Zephyr-7B-Beta, which delivered better output in terms of style and faithful content.

4.2 Human Evaluation

Experiment Design. To evaluate the three-phase search system in an end-to-end manner, we designed a user study in which participants explored publications related to two predefined search scenarios. They interacted with the conversational search interface and a graphical interface featuring a traditional text-based search, allowing us to compare the effectiveness of both systems. For the experiment, we recruited 40 participants from university courses and social networks according to criteria that match our target user group of non-experts. Table 3 in Appendix A gives an overview of participant demographics. All participants had at least basic technical knowledge, good English proficiency, and an interest in NLP without having expert-level knowledge. The gender composition was 35% female and 65% male, ranging in age from 20 to 29 years, with an average age of 25.

Prior to the user experiment, we randomly assigned participants into two groups (Group A and Group B), which determined the sequence in which each group used the search interfaces for the two search scenarios (Scenario 1 and Scenario 2). We ensured that the demographic characteristics of both groups were similarly distributed. Group A is exposed to Scenario 1 with the conversational interface first, followed by Scenario 2 with the graphical interface. Conversely, Group B is exposed to Scenario 1 with the graphical interface first, then Scenario 2 with the conversational interface. This

Evaluation Metric	**Scenario 1**		**Scenario 2**	
Mean (Std. Dev.)	**Conversational**	**Graphical**	**Conversational**	**Graphical**
System usability scale	76.00 (18.94)	77.25 (15.28)	76.63 (16.63)	65.25 (23.91)
Readability	4.50 (0.95)	3.40 (1.14)	4.45 (0.76)	3.20 (1.54)
Correctness	4.25 (0.97)	4.05 (1.00)	4.25 (0.72)	3.85 (1.31)
Usefulness	4.50 (0.61)	3.65 (0.99)	4.30 (0.80)	2.95 (1.23)
Summary quality	4.10 (0.85)	-	4.15 (0.67)	-
Overall satisfaction	4.15 (0.88)	3.45 (1.00)	4.10 (1.07)	2.85 (1.14)

Table 1: Overview of mean and standard deviation of evaluation metrics by search scenario and interface type.

crossover design allows each participant to test both interfaces but for a different scenario to avoid learning effects. Scenario 1 is about analyzing emotional expressions on social media related to mental health during the COVID-19 pandemic, while Scenario 2 focuses on creating multiple-choice exams for a programming course. Both scenarios end with the instruction: "Your task is to use the provided search interface below to find papers related to the described scenario." The full scenario descriptions are documented in Table 2 in Appendix A.

Participants were given approximately 10 minutes to interact with the interfaces (see screenshots in Table 4), followed by an evaluation questionnaire, which we share in the repository. The latter includes 10 questions from the *system usability scale* (SUS) (Brooke, 1996). The SUS metric is calculated using a specific formula, resulting in a value between 0 and 100, with 68 being considered average. Furthermore, participants were asked five questions to rate the general information quality in terms of readability, correctness, and usefulness, the quality of the generated comparisons, as well as overall satisfaction. The questions were answered on a 5-point Likert scale, where a rating of 5 denoted the most favorable value. In addition, we included two open-ended free-text fields for feedback on the system's strengths and weaknesses.

Quantitative Analysis of Evaluation Metrics. Based on the questionnaire responses, we conducted both quantitative and qualitative analyses. Table 1 lists the mean and standard deviation for each evaluation metric grouped by scenario and interface. We found that all data points were within reasonable ranges without containing significant outliers. Generally, the ratings for the conversational interface tend to be more favorable across the various evaluation metrics. This is also reflected in the overall satisfaction scores for Scenario 1 and Scenario 2, with ratings of 4.15 and 4.10 for the conversational interface compared to 3.45 and 2.85 for the graphical interface. The conversational interface especially surpasses the graphical interface in readability and usefulness metrics, as reflected by mean ratings that were around one point higher across both scenarios. We hypothesize that the dialogue interaction was not as overwhelming, delivering information in a more digestible format and increasing its overall utility by offering additional choices for paper selection. This positive feedback was likely also influenced by the quality of summarized paper comparisons, which achieved solid scores of 4.10 in Scenario 1 and 4.15 in Scenario 2.

Inspecting the perceived system usability, the SUS scores for Scenario 1 were similar for both interfaces, with SUS scores at 76.00 and 77.25. Since all participants presumably had prior experience on how to use a text-based search engine as well as a standard chat interface, it is not surprising that the scores are similar. It is very likely that each participant was already acquainted with operating both types of interfaces. In Scenario 2, the conversational interface maintained a comparable score of 76.63, whereas the graphical interface had a lower score of 65.25, suggesting that the conversational interface offers more consistent usability across different search scenarios. This observation is corroborated by the rating distributions illustrated in Figure 3. The latter shows that Scenario 2 received worse ratings compared to Scenario 1 for both evaluated interfaces, with the rating distributions shifting towards the lower scores, possibly indicating that Scenario 2 was slightly more difficult with regard to discovering relevant papers. This effect was especially pronounced for the graphical interface compared to the more consistent performance of the conversational interface. The more stable ratings of the conversational interface could suggest it retains usability and information relevancy, even during more challenging exploration tasks.

Qualitative Analysis of Participant Feedback. Our qualitative analysis of the free-text responses concerning the systems' strengths and weaknesses confirmed the quantitative results, revealing that

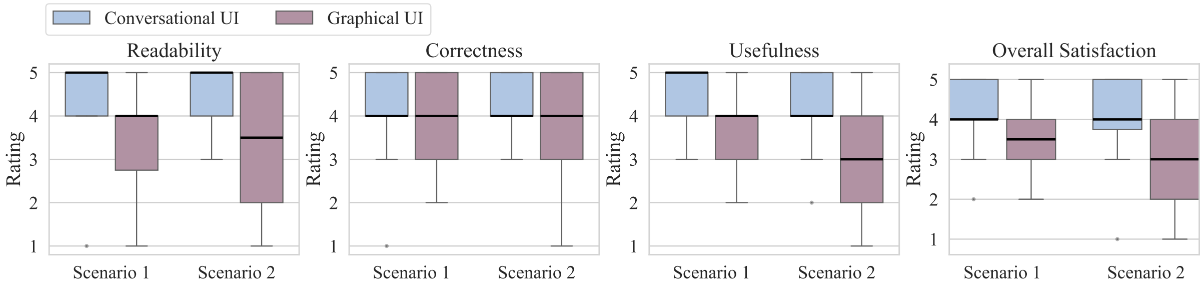


Figure 3: Comparison of rating distribution between the conversational and graphical user interface (UI) for readability, correctness, usefulness, and overall satisfaction. The thick line intersecting the box marks the median.

the conversational interface received higher ratings and demonstrated more consistent usability than the graphical interface. Every participant was assigned an anonymous identifier between P1 and P40 to protect their privacy. In the following paragraphs, cited questionnaire responses are presented in quotation marks and assigned to their identifiers.

The first notable strength of the conversational search interface, mentioned in nearly every second feedback comment, was the system's ease of use. For example, participant P9 remarked, "It is easy to use and to find topics & papers even if the prior exposure to the given topic is low." Users appreciated that they could immediately start talking with the conversational agent without requiring extensive knowledge of the interface or the search domain, unlike many graphical interfaces. A second strength highlighted by users was the system's guidance and structured navigation abilities, with one participant positively noting the "Step by step process to narrow down the search and avoid search results that are not related to your query" (P29). This feature effectively addressed the search-related vocabulary mismatches, as exemplified by the comment: "I don't need to know exact terms or what im looking for" (P40). Lastly, users valued the time-saving clustering and summarization features, which helped them avoid going through individual abstracts from long lists of papers. As participant P11 stated, "[...] it understands the content of the paper and can aggregate it, without me having to manually go into the files to read the Abstract." These findings suggest that conversational agents can help alleviate problems associated with cognitive overload (Sweller, 1988) by gradually communicating condensed information.

Yet, a couple of participants initially struggled with understanding the three-phase search flow (e.g., "In the first a few minutes it's hard to understand what I can reach at the end of conversation" (P35)). Some were also confused by the two options of selecting the suggested user response inputs displayed as buttons versus entering free-form text. This was especially the case when they wanted to reverse a choice, which participant P20 remarked, "The options to backtracking are a bit unclear at first." Other participants expressed a desire to "converse more freely" (P9), similar to those offered by general-purpose LLMs. Strengthening the integration of LLMs could accommodate this preference, as LLMs excel in contextual understanding of queries and navigating complex conversation logic more effectively than intent-based dialogue systems. We observed that certain users attempted to input very long requests or copy-pasting problem descriptions, an interaction more akin to LLM services like ChatGPT (OpenAI, 2022), where users input a prompt, check the output, and refine the prompt without engaging in a proper dialogue. This type of interaction does not align with how our task-oriented dialogue system was designed to operate. Nonetheless, the evaluation shows that nearly all participants quickly figured out how our conversational system works after a few dialogue turns.

5 Conclusion

We proposed a conversational exploratory search system integrated with a scholarly KG. Our study details the architectural components and presents results from training and evaluating language models that underpin the three-phase search process, including research topic classification, text clustering, and text summarization. We conducted a human evaluation to assess the system's effectiveness, identifying its perceived strengths and potential improvements. Our findings offer practical insights into the design and implementation of conversational search systems for the scholarly domain.

Acknowledgements

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

References

Simone Angioni, Angelo Salatino, Francesco Osborne, Diego Reforgiato Recupero, and Enrico Motta. 2021. Aida: A knowledge graph about research dynamics in academia and industry. *Quantitative Science Studies*, 2(4):1356–1398.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP*, pages 3615–3620, Hong Kong, China. ACL.

Lutz Bornmann, Robin Haunschild, and Ruediger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8.

John B. Brooke. 1996. Sus: A 'quick and dirty' usability scale. *Usability evaluation in industry*, 189(194):4–7.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Conference on EMNLP and the IJCNLP*, pages 1803–1813, Hong Kong, China. ACL.

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 EMNLP: Industry Track*, pages 811–819, Singapore. ACL.

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*, pages 729–738.

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*.

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Sérgio Gonçalves, Paulo Cortez, and Sérgio Moro. 2020. A deep learning classifier for sentence classification in biomedical and computer science abstracts. *Neural Computing and Applications*, 32(11):6793–6807.

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).

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint*.

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*.

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. *NeurIPS*, 33:9459–9474.

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *European Semantic Web Conference*, pages 182–185. Springer.

Antonello Meloni, Simone Angioni, Angelo Salatino, Francesco Osborne, Aliaksandr Birukou, Diego Reforgiato Recupero, and Enrico Motta. 2023a. Aidabot 2.0: Enhancing conversational agents with knowledge graphs for analysing the research landscape. In *International Semantic Web Conference*, pages 400–418. Springer.

Antonello Meloni, Simone Angioni, Angelo Salatino, Francesco Osborne, Diego Reforgiato Recupero, and Enrico Motta. 2023b. Integrating conversational agents and knowledge graphs within the scholarly domain. *IEEE Access*, 11:22468–22489.

Fionn Murtagh and Pedro Contreras. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97.

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

Enrique Orduña Malea, Alberto Martín-Martín, Juan M Ayllón, and Emilio Delgado-López-Cózar. 2014. The silent fading of an academic search engine: the case of microsoft academic search. *Online Information Review*, 38(7):936–953.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on EMNLP*, pages 11670–11688, Abu Dhabi, United Arab Emirates. ACL.

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. ACM.

Phillip Schneider, Anum Afzal, Juraj Vladika, Daniel Braun, and Florian Matthes. 2023a. Investigating conversational search behavior for domain exploration. In *Proceedings of the 45th European Conference on Information Retrieval (ECIR)*, pages 608–616, Dublin, Ireland. 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 *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 807–814, Rome, Italy. SciTePress.

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 (EACL)*, pages 358–367, St. Julian's, Malta. Association for Computational Linguistics.

Phillip Schneider, Wessel Poelman, Michael Rovatsos, and Florian Matthes. 2024c. Engineering conversational search systems: A review of applications, architectures, and functional components. In *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI)*, pages 73–88, Bangkok, Thailand. 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 (PACLIC)*, pages 609–619, Hong Kong, China. Association for Computational Linguistics.

Tim Schopf, Karim Arabi, and Florian Matthes. 2023. Exploring the landscape of natural language processing research. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1034–1045, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tim Schopf and Florian Matthes. 2024. Nlp-kg: A system for exploratory search of scientific literature in natural language processing. In *Proceedings of the 62st Annual Meeting of the ACL (Volume 3: System Demonstrations)*, Bangkok, Thailand. ACL.

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. ACM.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on EMNLP*, pages 5548–5566, Singapore. ACL.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. In *NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing*.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020a. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*, volume 33, pages 5776–5788.

A Appendix

The Appendix provides further insights into the results of our research, including a finite-state diagram of the conversational search flow (Figure 4), the semantic data model of the scholarly KG (Figure 5), the scenario descriptions for the human evaluation (Table 2), a tabular overview of the participant demographics (Table 3), screenshots of the conversational and graphical interface (Table 4), and a collection of the crafted LLM prompts (Table 5).

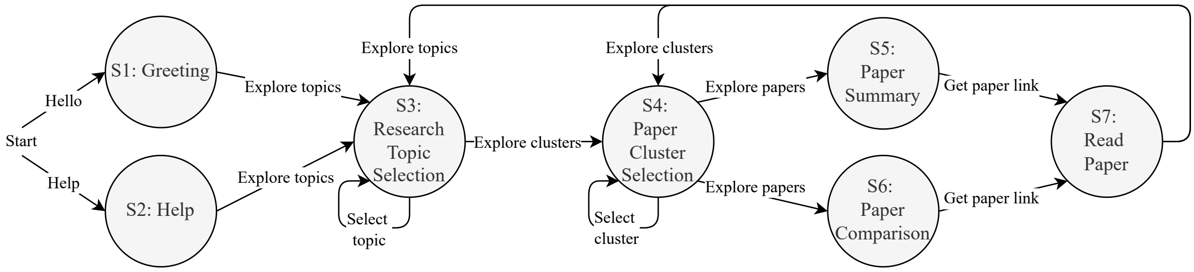


Figure 4: Conversational search flow illustrated as dialogue states (S1-S7). The three-phase search process encompasses: first, identifying a research topic (S3); second, choosing clusters of publications (S4); and third, comparing publications via short summaries (S5-S6).

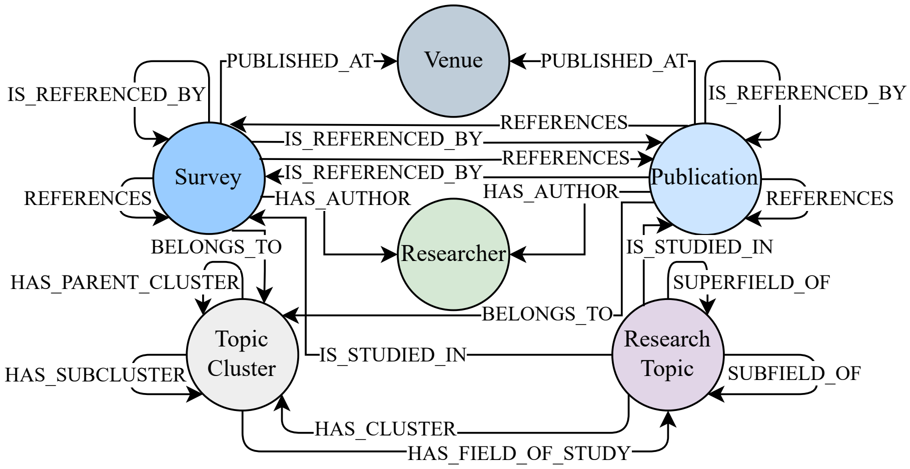


Figure 5: Semantic data model of the scholarly knowledge graph.

Description of Scenario 1
Imagine you are interested in mental health and emotional changes during the COVID pandemic. You want to analyze how people express their feelings on social media platforms during the pandemic. Your goal is to study their emotions to learn how they handle stress and anxiety. To deal with the enormous volume of data available online, you are looking for ways to automate the analysis process using NLP techniques. Your task is to use the provided search interface below to discover papers related to the described scenario. You have up to 8 minutes for your exploratory search. You are encouraged to "think out loud". Afterward, you will fill out an evaluation questionnaire to provide feedback on your search experience.
Description of Scenario 1
Imagine you are an instructor teaching a programming course in Python. You want to ensure that the exam questions reflect the learning progress of the students. Your goal is to generate multiple-choice exams according to their knowledge level. To achieve this, you are looking for ways to automatically create exam questions based on the course materials using NLP techniques. Your task is to use the provided search interface below to discover papers related to the described scenario. You have up to 8 minutes for your exploratory search. You are encouraged to "think out loud". Afterward, you will fill out an evaluation questionnaire to provide feedback on your search experience.

Table 2: Scenario descriptions of the exploratory search task for the human evaluation.

Demographic Variable	**Group A (n = 20)**	**Group B (n = 20)**	**Overall (n = 40)**
Mean age (age range)	25.10 (20 to 29)	24.95 (23 to 28)	25.03 (20 to 29)
Male	13	13	26
Female	7	7	14
High school degree	1	-	1
Bachelor's degree	14	17	31
Master's degree	5	3	8
No NLP knowledge	1	2	3
Beginner NLP knowledge	15	13	28
Advanced NLP knowledge	4	5	9
English CEFR level B1 or B2	3	5	8
English CEFR level C1 or C2	17	15	32

Table 3: Overview of study participant demographics.

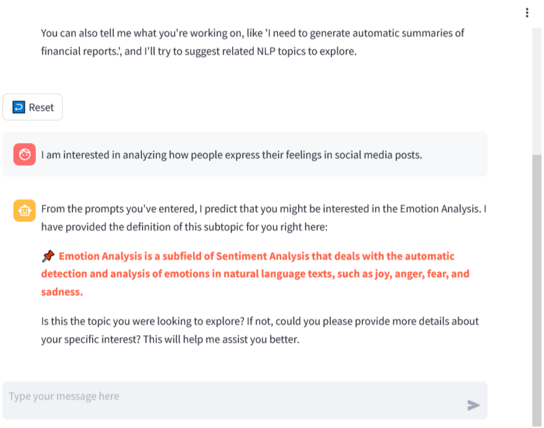
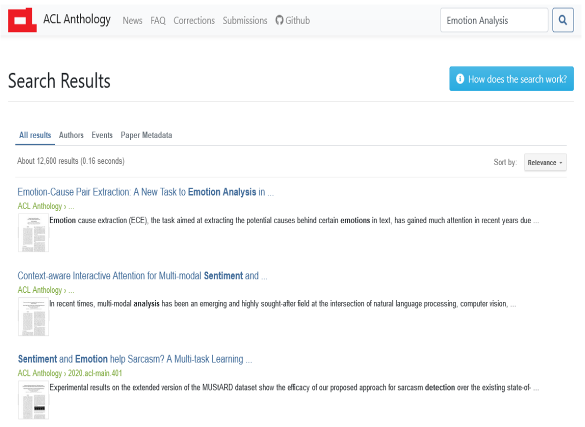
Conversational Interface	**Graphical Interface**
 You can also tell me what you're working on, like 'I need to generate automatic summaries of financial reports.', and I'll try to suggest related NLP topics to explore. I am interested in analyzing how people express their feelings in social media posts. From the prompts you've entered, I predict that you might be interested in the Emotion Analysis. I have provided the definition of this subtopic for you right here: 📌 **Emotion Analysis is a subfield of Sentiment Analysis that deals with the automatic detection and analysis of emotions in natural language texts, such as joy, anger, fear, and sadness.** Is this the topic you were looking to explore? If not, could you please provide more details about your specific interest? This will help me assist you better. 	 ACL Anthology News FAQ Corrections Submissions Github Emotion Analysis Search Results How does the search work? All results Authors Events Paper Metadata About 12,600 results (0.16 seconds) Sort by: Relevance Emotion-Cause Pair Extraction: A New Task to **Emotion Analysis** in ... ACL Anthology › ... **Emotion** cause extraction (ECE), the task aimed at extracting the potential causes behind certain **emotions** in text, has gained much attention in recent years due ... Context-aware Interactive Attention for Multi-modal **Sentiment** and ... ACL Anthology › ... In recent times, multi-modal **analysis** has been an emerging and highly sought-after field at the intersection of natural language processing, computer vision, ... **Sentiment** and **Emotion** help Sarcasm? A Multi-task Learning ... ACL Anthology › 2020.acl-main.401 Experimental results on the extended version of the MUStARD dataset show the efficacy of our proposed approach for sarcasm **detection** over the existing state-of- ...

Table 4: Visual side-by-side comparison of the conversational and graphical interface from the human evaluation.

Prompt 1: Cluster Name Generation (Zephyr-7B-Beta)

Considering the themes and topics from the following TFIDF cluster tag: "{tfidf_cluster_name}", please provide a concise and descriptive name for a cluster that includes these {len(paper_list)} academic papers:
{paper_titles_formatted}
Respond with just the cluster name, based on the overarching themes evident in the titles and the TFIDF tag. Don't include the original TFIDF cluster tag and the word "Cluster" in your response.

Prompt 2: Comparative Text Summarization (Zephyr-7B-Beta)

Please provide a comparative analysis of the objectives of two scientific papers.
Refer the papers with their real ids:
Paper {id_a}'s objective is: {obj1}
Paper {id_b}'s objective is: {obj2}
Highlight the key differences and similarities between Paper {id_a} and Paper {id_b}.
Use simple language.
Please provide a comparative analysis of the results of two scientific papers.:
Refer the papers with their real ids:
Results of Paper {id_a}: {res1}
Results of Paper {id_b}: {res2}
Highlight the key differences and similarities between Paper {id_a} and Paper {id_b}.
Use simple language.
Please provide a comparative analysis of the TLDR of two scientific papers.:
TLDR of Paper {id_a}: {tldr1}
TLDR of Paper {id_b}: {tldr2}
Highlight the key differences and similarities between Paper {id_a} and Paper {id_b}.
Use simple language.

Prompt 3: LLM-Based Research Topic Classification (GPT-3.5-Turbo)

You are supposed to classify a query into one of the topics provided. These topics are various fields of NLP. Your answer should be in the following format:
topic name
Nothing else should be included in the output.
Make sure there is no extra punctuation including full stops, quotation marks or anything of that sort. You are supposed to EXACTLY use the topics from the list provided. If you think it is a random question and not in the field of NLP, then return the topic as "none".
You can only provide your answer from the following topics and the topics are:
Multimodality
Natural Language Interfaces
Semantic Text Processing
Semantic Analysis
Syntactic Text Processing
Linguistic and Cognitive NLP
Responsible NLP
Reasoning
Multilinguality
Information Retrieval
Information Extraction and Text Mining
Text Generation
Query: {query}.
Topic:

Table 5: Overview of large language model prompts for various generative tasks using Zephyr-7B-Beta and GPT-3.5-Turbo. Dynamically inserted variables are enclosed within curly brackets.