ACLBot: A Knowledge Graph-Driven Assistant for ACL Anthology Research

Jan Buchmann[1*], **Steven Lynden**[2], **Kristiina Jokinen**[2,3]
[1] Ubiquitous Knowledge Processing Lab
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
www.ukp.tu-darmstadt.de
[2] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[3] University of Helsinki, Finland
jan.buchmann@tu-darmstadt.de
steven.lynden@aist.go.jp, kristiina.jokinen@aist.go.jp

Abstract

We present ACLBot, an interactive chatbot designed to support literature exploration in the ACL Anthology by combining structured knowledge graph querying with large language model (LLM) generative AI. ACLBot integrates a Neo4j-based knowledge graph constructed by extracting data on publications, authors, topics, and research trends from the ACL Anthology, and automatically generates knowledge graph queries to retrieve relevant information in response to user questions. Retrieved results are re-injected into the LLM to produce concise, contextually grounded summaries. We describe the system's architecture, including its query generation pipeline, knowledge graph integration, and visualization components for highlighting temporal trends in research. To assess usability and effectiveness, we conducted a user evaluation with researchers, collecting qualitative and quantitative feedback on response accuracy, informativeness, and utility for literature discovery. Results indicate that ACLBot effectively supports exploratory search, helps identify relevant works and trends, and offers a promising framework for integrating structured information with generative AI for scientific information retrieval.

Keywords: Information Extraction, Knowledge Graphs, Dialogue Systems

1. Introduction

Exploring the enormous volume of the ACL Anthology (Bird et al., 2008) can be a daunting task for researchers, who must sift through thousands of papers to locate relevant work, authors, and trends. Although state-of-the-art large language models (LLMs) such as GPT-5 are capable of assisting in such tasks, they lack detailed knowledge about specific areas and are prone to hallucination. To address this, we introduce **ACLBot**, an interactive system that blends structured retrieval from a knowledge graph with the generative power of LLMs.

Knowledge graphs, structured, graph-based representations of entities and their relationships, have become powerful tools in AI for semantic understanding, reasoning, and explainability across domains (Hogan et al., 2021). In scholarly search, graph-based systems have already shown promise in supporting exploratory and conversational interfaces for literature discovery (Christen et al., 2023). Meanwhile, graph databases like Neo4j (Eifrem et al., 2010), when paired with Cypher (Francis et al., 2018) query capabilities, can offer expressive and efficient access to interlinked scholarly metadata (Neo4j Inc., 2023).

ACLBot leverages this synergy by mapping user questions to Cypher queries, executing them over a Neo4j-based knowledge graph, and then re-injecting the retrieved results into an LLM to produce concise and grounded summaries. This hybrid approach mitigates hallucination risks while supporting flexible, exploratory conversation.

In addition to presenting ACLBot's architecture encompassing Cypher generation, knowledge graph integration, and visualization components, we report the results of a user evaluation assessing accuracy, informativeness, and usability. Overall, our contribution lies in demonstrating how symbolic–neural integration enables more effective literature exploration in NLP-focused corpora.

The remainder of this paper is structured as follows. Section 2 reviews prior work on knowledge graphs, conversational agents, and scholarly information retrieval. Section 3 details the construction of our ACL Anthology knowledge graph, including data sources, extraction methods, and schema design, followed by Section 4 which assesses data quality. Section 5 presents the architecture of ACLBot, describing its interaction layers and chart generation capabilities. Section 6 outlines the user evaluation methodology and reports the results. Section 7 discusses the limitations of our current approach and directions for future research. Finally, Section 8 concludes the paper.

*Work conducted while the author was a technical trainee at AIST.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2731–2741
11–16 May 2026.

2. Related Work

Our work builds upon prior research at the intersection of knowledge graphs, conversational agents, and scholarly information retrieval.

Knowledge graphs (KGs) in scholarly search. KGs have emerged as a central paradigm for representing structured knowledge and supporting semantic search (Hogan et al., 2021). In the scholarly domain, KGs such as the Microsoft Academic Graph (Sinha et al., 2015) and Open Research Knowledge Graph (Auer et al., 2020) have been used to organize metadata, support advanced querying, and enable semantic discovery.

Conversational interfaces for KGs. Recent studies have explored conversational agents capable of interpreting natural language queries over KGs (Schneider et al., 2023; Christen et al., 2023; Shen et al., 2020). Such systems translate user utterances into formal queries (e.g. SPARQL, Harris and Seaborne 2013, or Cypher) and present results in an interactive dialogue, a paradigm that inspires the design of ACLBot.

LLM–KG integration. Several works integrate large language models with structured data to improve factual grounding and reduce hallucinations (Pan et al., 2023; Agarwal et al., 2023). These approaches combine the reasoning capabilities of LLMs with the precision of KGs, often through retrieval-augmented generation (Lewis et al., 2020) (RAG) pipelines. ACLBot follows this hybrid architecture, tailored to the ACL Anthology corpus.

Evaluation of chatbot assistants. Prior research has surveyed evaluation methods for chatbots, highlighting dimensions such as accuracy, user satisfaction, and efficiency (Maroengsit et al., 2019). We adopt similar metrics, focusing on ACLBot's support for exploratory search and trend analysis in the ACL Anthology.

3. Knowledge Graph Construction

Our knowledge graph (KG) is implemented in Neo4j and built from the ACL Anthology corpus, enriched with structured metadata, extracted research entities, experimental results, and dense text embeddings. The resulting KG serves as the structured foundation for ACLBot's hybrid retrieval and grounded response generation.

Corpus preprocessing. We obtained the information in the knowledge graph from two sources: The ACL anthology package[1] for metadata such as authors, venues, and abstracts, and paper PDFs for information from full texts. As extracting information directly from PDFs is difficult, we converted these files to Intertext Graph (ITG)[2] and plain text. ITG encodes document structure (titles, sections, paragraphs) and bibliographic metadata. We obtained ITG files by first converting PDFs to XML with GROBID (GRO, 2008–2025)[3], and then parsing to ITG. GROBID often fails to recognize tables correctly, which hinders result extraction. For this purpose, we used pypdf to convert PDFs to plain text with improved table representation, but missing document structure. Where abstracts were missing in the ACL anthology package, we re-extracted them directly from ITG segmentations.

LLM-based extraction. To populate the KG with fine-grained research information, we employed a GPT-4-class large language model with domain-specific prompts. The extraction pipeline generated:

1. **Research classifications:** each paper was assigned one `Area` (from a fixed set of 20) and one or more `ContributionType` labels (from 13 categories), which were taken from the ACL ARR call for papers.[4]
2. **Entities:** mentions of `Task`, `Dataset`, `Metric`, `Architecture`, `Method`, and `PretrainedModel` were extracted.
3. **Results:** As in related work on automatic leaderboard construction (Şahinuç et al., 2024), experimental results were represented as `Result` nodes, each linked to a task, dataset, and metric, storing the reported score.

Extraction prompts enforced JSON schemas, ensuring structured outputs that could be reliably ingested. Intermediate results were stored in JSON before ingestion into Neo4j.

Graph schema. The KG schema covers publication metadata (`Paper`, `Author`, `Event`, `Volume`, `Passage`), research classifications (`Area`, `ContributionType`), research entities, and evaluation outcomes (`Result`). Edges represent authorship (`AUTHORED_BY`), publication venue (`PUBLISHED_IN`, `BELONGS_TO`), document structure (`CONTAINS`, `IS_PARENT_OF`,

[1] https://github.com/acl-org/acl-anthology

[2] https://github.com/UKPLab/intertext-graph

[3] Where available, we used existing XMLs from the ACL Anthology Corpus (Rohatgi, 2022).

[4] https://aclrollingreview.org/cfp

Node label	Count	Rel. type	Count
Passage	12,110,037	CONTAINS	12,110,037
Result	238,678	IS_FOLLOWED_BY	6,198,030
Paper	105,626	IS_PARENT_OF	6,198,030
Dataset	97,303	ON	716,034
Author	95,444	WORKS_ON	443,260
Method	63,245	AUTHORED_BY	361,414
Metric	53,115	USES	328,484
Task	39,674	REPORTS	238,486
Architecture	31,940	HAS_CONTRIBUTIONTYPE	142,187
PretrainedModel	11,160	PUBLISHED_IN	105,626
Volume	3,105	HAS_AREA	76,832
Event	1,922	BELONGS_TO	3,105
Area	25		
ContributionType	13		

Table 1: Node and relationship counts in the ACLBot knowledge graph.

`IS_FOLLOWED_BY`), semantic associations (`HAS_AREA`, `HAS_CONTRIBUTIONTYPE`, `USES`, `WORKS_ON`), and evaluation results (`REPORTS`, `ON`).

Passages and embeddings. Each paper was segmented into `Passage` nodes (document titles, section titles, paragraphs). To enable semantic search, passages and abstracts were embedded using the `all-MiniLM-L6-v2` model (384-dimensional sentence embeddings). These vectors are stored as node properties, enabling approximate nearest-neighbor search via Neo4j's vector index. This allows ACLBot to perform semantic retrieval when symbolic Cypher queries are insufficiently specific.

Indexing and optimization. We created uniqueness constraints on identifiers (e.g., `Paper.id`, `Dataset.name`), full-text indexes on textual fields (titles, abstracts, passages), and vector indexes on embedding properties. The indexing supports both symbolic retrieval (via Cypher) and dense semantic retrieval (via vector similarity search).

Database statistics. The resulting KG contains tens of thousands of papers, each enriched with bibliographic metadata, research classifications, extracted entities, and experimental results. The total size on disk is 74.6GB and the number of the various node types and relationship types is shown in Table 1. It supports dual retrieval modes: (i) symbolic queries for precise filtering and relation traversal, and (ii) neural semantic search for approximate matching over abstracts and passages. This hybrid retrieval capability underpins ACLBot's ability to generate factually grounded answers.

4. Data Quality Assessment

To estimate the reliability of the data in the knowledge graph (§3), we performed an evaluation of the data extraction. For `Area`, `ContributionType` and research entities (`Task`, `Model`, etc.), one author manually reviewed 100 extracted items per category and annotated their correctness. Table 3 shows that precision is above 0.8 for all categories, suggesting that extraction is reliable. Due to time constraints, we did not judge the recall of extraction, which takes significantly more effort. To estimate the reliability of `Result` extraction, we ran GPT-4o-mini on the SciLead dataset (Şahinuç et al. 2024). Precision and recall of full result extraction are at 0.52, while for results without scores, they are at 0.71 and 0.62, respectively (Table 4). This shows that result extraction is less reliable, and thus ACLBot answers about results should be cross-checked.

5. System Overview

In this section the core functionality of the ACLBot [5] system is described. Figure 1 illustrates the overall flow of ACLBot's operation. A user's natural language prompt is received and Microsoft's Semantic Kernel [6] orchestrates the routing of the request. Depending on the intent, ACLBot either translates the prompt into a Cypher query for symbolic retrieval over the knowledge graph or invokes semantic passage search using vector embeddings.

Application Layer. At the core is Semantic Kernel, which provides an orchestration framework

[5] Source code available at https://github.com/nsjl/aclbot

[6] https://learn.microsoft.com/en-us/semantic-kernel/overview/

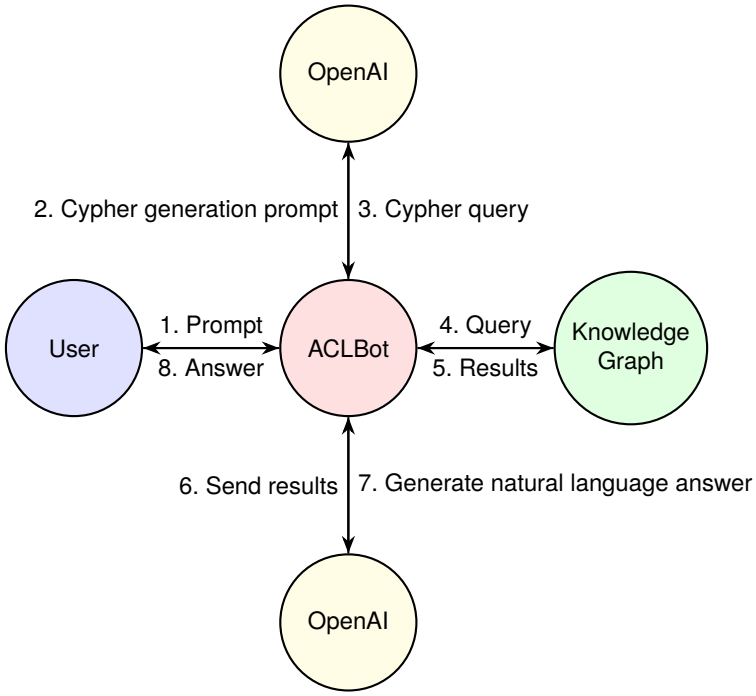


Figure 1: ACLBot system flow: the user enters a prompt (1). If the request is symbolic, a Cypher query is generated (2–3), executed over the Neo4j KG (4), and results are returned (5). These results are then sent back to the LLM (6–7), producing the final answer (8). For semantic requests, the system directly retrieves passages using vector search before steps 6–7.

connecting large language models with symbolic and neural retrieval. When a prompt maps naturally to structured graph queries (e.g., "list all datasets used with BERT"), Semantic Kernel formulates a Cypher statement via an LLM and executes it against the knowledge graph (Steps 2–5 in Figure 1). When the query is exploratory or meaning-based (e.g., "papers about efficient translation models"), the system instead retrieves semantically similar passages using vector indexes over text embeddings. In both cases, retrieved results are forwarded for answer generation.

Data Access Layer. This layer interfaces with the knowledge graph described in Section 3. Symbolic retrieval operates over the graph's structured schema (papers, authors, tasks, datasets, results), while semantic retrieval leverages embeddings stored on passage and abstract nodes, enabling approximate nearest-neighbor search.

Response Generation. Retrieved results are passed to the LLM for natural language generation (Step 7). This stage fuses factual content from the KG or semantic search with coherent textual phrasing before delivering the final answer to the user (Step 8).

Hybrid Retrieval. By combining symbolic (Cypher-based) and neural (embedding-based) retrieval paths, ACLBot supports both precise fact-finding and broader exploratory search. The Semantic Kernel framework provides a unified control layer, allowing the system to dynamically decide which retrieval mode to apply, and ensuring that both paths integrate seamlessly into the response pipeline.

Chart Generation. To complement textual answers, ACLBot can opportunistically generate figures (bar, pie, scatter, or heatmap) summarizing entity usage. This lightweight module inspects retrieved results and user prompts for topical cues, and where applicable, renders compact charts using `matplotlib` [7] and `seaborn` [8]. These figures provide quick visual summaries for common entity-centric queries without introducing latency into the main pipeline. Figure 2 shows an example output.

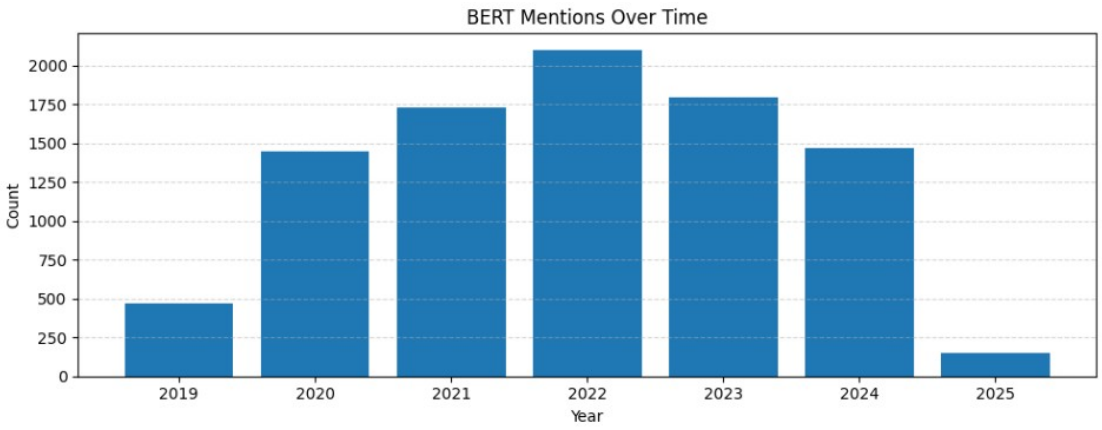
6. User Evaluation

We conducted a user evaluation to assess the usefulness, clarity, and accuracy of the system's outputs. Due to organizational restrictions on releasing a live Web-accessible deployment of ACLBot and the difficulty in requiring installation of the approximately 80GB knowledge graph, a dedicated evalua-

[7] https://matplotlib.org/
[8] https://seaborn.pydata.org/

User: What does the trend in BERT adoption look like?

Bot:



Bot: The trend in BERT adoption over the years, based on the number of papers that mention it, is as follows:

- In 2010 and 2015, there was minimal mention of BERT in research papers, with only 1 paper each year.
- It first appeared in 2018 with 1 paper, which marked the beginning of its adoption.
- In 2019, BERT was mentioned in 439 papers, indicating a significant increase in interest.
- The number of papers continued to rise in 2020 with 1225 papers and peaked in 2022 with 1485 papers.
- In 2023, the number dropped slightly to 1233 papers.
- Predictions for future years show a decline in mention, with 941 papers in 2024 and just 84 papers in 2025.

This trend shows a rapid rise in the adoption of BERT after its introduction, peaking around 2022, and then a gradual decline in its mention in subsequent years.

Figure 2: Example ACLBot answer combining figure with textual reply.

tion Web portal was provided presenting evaluators with a list of six of pre-recorded system–user conversations:

- **Model and Metrics Trends over Time** This chat examines model and metric trends utilized by ACL anthology research papers in the ACLBot database.
- **Research Practices, Methods and Metrics** Questions and answers about the usage of metrics, methods etc. in the knowledge graph.
- **Adoption of BERT** On the emergence of papers on or using BERT circa 2018. (Part of this conversation is visualized in Figure 2.)
- **What's in the Database?** Chat about what is in the ACLBot's database.
- **Random Paper** Ask questions about a specific paper selected at random.
- **BERTScore** Chat about the emergence of the BERTScore metric.

Participants could browse and select any conversation of interest, after which they were redirected to the evaluation form to submit their assessment. The evaluation process was implemented via a Web-based interface designed for ease of access and anonymous participation. We used Google Forms to collect relevant information from each evaluator as follows:

1. **Information about the evaluator**
   Evaluator position [student, postdoc, faculty, or other] and familiarity with NLP and the ACL Anthology research [1–5 scale].

2. **Evaluation of chat contents**
   A set of 5-point Likert-scale statements concerning various aspects of ACLBot:

Category/entity	Count
Top 5 Areas	
Information Extraction	10,416
NLP Applications	9,227
Machine Translation	7,765
Efficient/Low-Resource Methods for NLP	4,784
Dialogue and Interactive Systems	4,176
Top 5 ContributionTypes	
NLP engineering experiment	69,688
Data resources	26,229
Data analysis	18,535
Publicly avail. software/pre-trained models	13,006
Approaches to low-resource settings	5,593
Top 5 Methods	
Expectation-Maximization	1,158
Logistic Regression	1,143
Data Augmentation	1,077
Fine-tuning	1,013
Contrastive Learning	839
Top 5 Architectures	
Transformer	10,539
BERT	9,158
RoBERTa	3,693
LSTM	3,201
T5	2,048
Top 5 PretrainedModels	
BERT	6,785
GloVe	2,477
RoBERTa	1,711
GPT-4	1,205
word2vec	1,020
Top 5 Tasks	
Machine Translation	3,433
Named Entity Recognition	3,018
Sentiment Analysis	2,514
Question Answering	1,922
Natural Language Inference	1,378
Top 5 Datasets	
CoNLL-2003	3,180
Europarl	935
Penn Treebank	688
Wikipedia	661
GLUE	644
Top 5 Metrics	
Accuracy	33,378
F1-score	27,888
BLEU	11,551
ROUGE-L	4,928
ROUGE-1	3,334

Table 2: Top 5 entries by category in the knowledge graph.

	Precision
`Area`	0.87
`ContributionType`	0.81
Entities	0.90

Table 3: Human evaluation of precision of `Area`, `ContributionType` and entity extraction.

P	R	F1
Full Results		
0.52	0.52	0.52
Without Score		
0.71	0.62	0.66

Table 4: Evaluation of result extraction on the SciLead dataset (Şahinuç et al., 2024).

- *Accuracy and relevance:* ACLBot answers were accurate/ relevant.
- *Usefulness and rationality:* ACLBot answers were useful/rational and sensible.
- *Comparative value:* ACLBot is a useful tool to help explore academic papers
  ACLBot can provide additional value over standard LLM-based chatbots such as GPT-4.
- *Error detection:* Did the assistant make any obvious factual errors? [Yes/No]

3. **Additional information**
   Open-ended, short paragraph-based entry allowing asking the evaluators: What did you like about the chat, or the idea of ACLBot in general?; What did you dislike about the chat, or the idea of ACLBot in general?; What is your overall view of the system? Would you recommend it?; Do you have any ideas on how ACLBot could be improved? Any features you would add?

4. **Optional questions**
   All prior questions were mandatory. Two optional questions were added as follows: If you are familiar with the concept of knowledge graphs, did ACLBot's answers show evidence of effective utilization of a knowledge graph? [yes/no/do not know] and the opportunity to add any additional comments (free text entry).

Responses were collected anonymously to encourage honest and unbiased feedback. Each evaluator could assess multiple conversations, and no limits were placed on the number of evaluations

submitted per participant. The collected feedback was subsequently aggregated and analyzed to identify recurring strengths and weaknesses of the system, the results of which are discussed below.

6.1. Results and Discussion

6.1.1. Evaluators

Participants were recruited from an HCI class, selected slack group, and staff members in the authors' institutions. They were identified as PhD students (61.5%), faculty members (17.9%), postdoctoral researchers (15.4%), and others (5.1%). In terms of age, most respondents were in the 21–30 age range (66.7%), with additional representation from those aged 31–40 (30.8%) and a small proportion over 60 (2.6%). The median familiarity score for natural language processing and LLMs was 5, while familiarity with the ACL Anthology had a median of 4. We can therefore conclude that responders were relatively familiar with the domain.

6.1.2. Responses

Due to the anonymity process involved in collecting the results, the total number of individuals who responded is not recorded, however 39 responses were received, fairly evenly distributed across the chats as follows:

- Model and Metric Usage Trends : 9 responses
- Adoption of BERT : 6 responses
- Explore a Random Paper : 6 responses
- BERT Score : 6 responses
- Research Practices and Methods : 6 responses
- What is in the Database : 6 responses

Qualitative evaluations of individual chats are displayed in Figure 3 and overall answers for the bot in Figure 4. The accuracy, relevance, usefulness and rationality of the ACLBot answers were rated highly by the evaluators. It is interesting that ACLBot was judged high on relevance (answers dealt with the correct topic), but not high on the content of the answers (rationality and sensibility). Based on the full chat-specific breakdown of results, the two metric/results-related chats ("Model and Metric Usage Trends", "BERT Score") averaged 3.47 for accuracy, whereas other chats averaged 3.38, which demonstrates that users observed the less-accurate extraction quality of metrics, especially their numerical values, as discussed in Section 4 and free-form comments in Section 6.1.3.

Figure 5 shows answer distribution to the two optional questions. Firstly, it is important to note that only 36% of responses agreed with the statement that ACLBot made no factual errors, indicating that as with many generative AI approaches there was significant hallucination in the generated answers. Secondly, almost 68% of responses were positive to the question of whether evidence of an underlying KG was present in answers, indicating observable benefits of our structured KG-approach to support generative models.

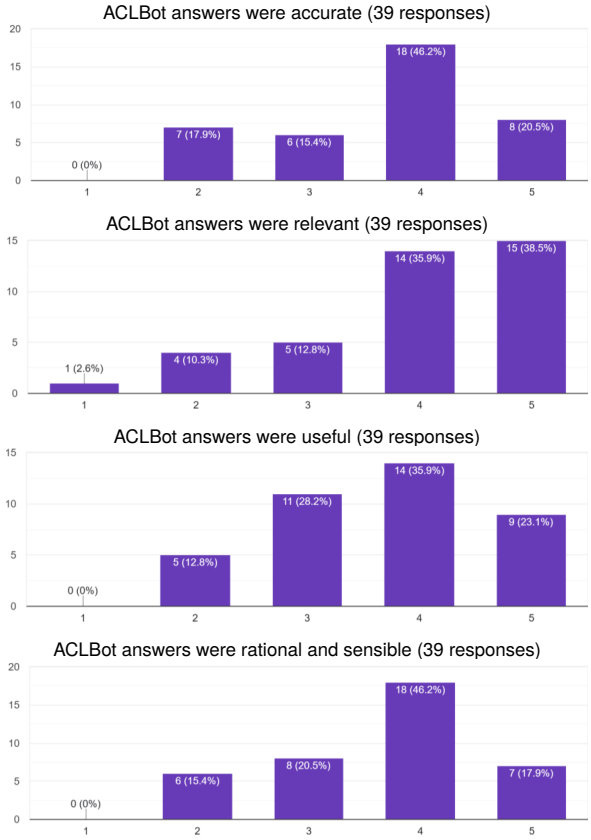


Figure 3: Chat-specific answers about accuracy, relevance, usefulness, and whether answers were sensible.

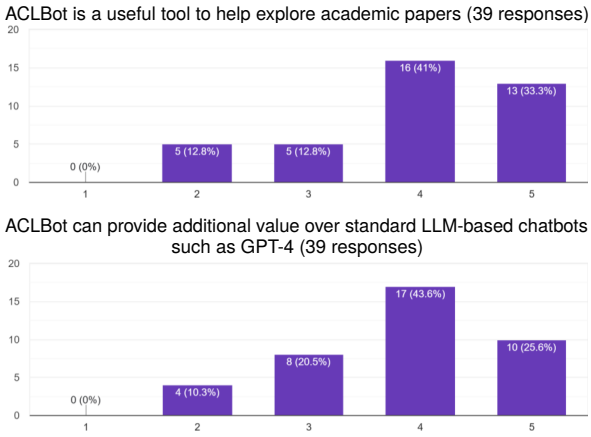


Figure 4: General (non chat-specific) answers from users about ACLBot.

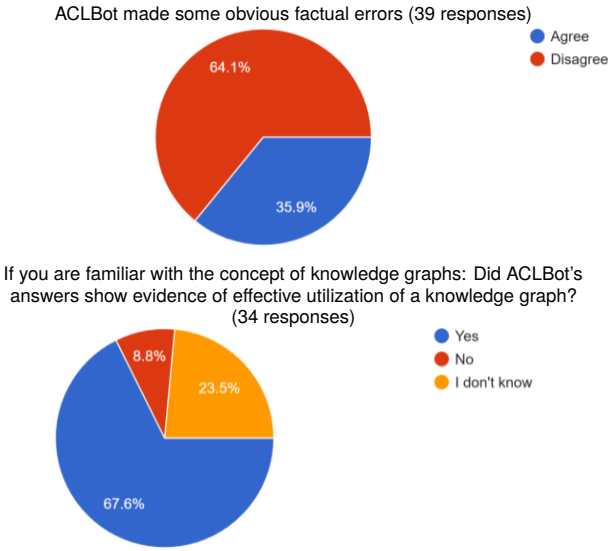


Figure 5: Answers to the optional questions: factual errors made by ACLBot and impressions on the evidence of knowledge graph usage by ACLBot.

6.1.3. Analysis of open-ended comments

The respondents provided positive feedback, but also pointed out issues to be improved for the future system. The comments show appreciation especially towards the interactive nature of search which enabled longer dialogues *"long dialogues with interactive search is good,"*(25%) and helpful summaries and visualization *"the fact that it could create relevant graphs and statistics"* (25%). Most also complimented the retrieved information being relevant and emphasized the value of the system's analytic capabilities, such as *"its ability to list papers based on metrics"* and *"finding the correct paper as evidence for utility."* These remarks suggest that users perceived ACLBot as an engaging and innovative tool for exploring academic literature and identifying patterns in research data.

On the other hand, although 14% of the responses did not report anything they disliked in the system, there were many comments (39%) on inconsistencies *"sometimes graphs were produced, sometimes not"* or *"It was weird that it first claimed there's a "Translation" paper, but didn't fetch any results about it"*, and inaccuracies involving numerical results or evaluation metrics, where answers were sometimes perceived as less precise *"suggested papers only from the same year"*. Two responses also provided observations on the phrasing of the answers which expressed uncertainty but were considered inappropriate in the given context *"Last question is also answered like "they might use these metrics", this is not an actual comparison"* or *"As ACL anthology may lack metadata for early papers, it assumes knowledge. Phrases like "assume", "probably", "seemingly" indicate this"*.

The users also provided comments on usability issues, which are important considering development of AI-based systems in general. For instance, they emphasised the need for good NLU components (*"It felt that the system did not understand the intent behind the query."*) as well as accurate response generation. Unreliable numerical information can severely hurt the user's trust in the system's functionality *"The numbers weren't reliable. Once I noticed it, I started to doubt all other answers"*.

Despite the critical comments, 46% of the responses gave unwavering recommendation for the system, and only 18% expressed strong unwillingness to do so. The rest (36%) acknowledged the benefits as well as pointed out various challenges.

Suggestions for future development included making interaction more natural (the system should give more feedback and also be allowed to ask clarifications) and enhancing the quality of numerical counting. Linking of the answers to ACL Anthology for easy verification was mentioned in 25% of the responses, and some suggested visualisation as a mandatory rather than an optional feature.

7. Limitations and Future Work

A key limitation of the current ACLBot is its reliance on prompt-based interactions with the underlying GPT model. Small modifications to the system prompt, intended to improve performance, can inadvertently degrade performance. Similarly, minor variations in the wording of user prompts may yield noticeably different outputs, reflecting the sensitivity of large language models to input phrasing.

Accuracy of metrics and their numerical scores was shown to be less reliable than other extracted data, and state-of-the-art approaches have demonstrated this to be a challenging active research problem (Şahinuç et al., 2024). We plan to take advantage of ongoing developments in this area in future versions of ACLBot.

In addition, ACLBot inherits well-known limitations of large language models related to hallucinations and overgeneralization. Although the system is designed to ground responses in a structured knowledge graph and retrieved text passages, erroneous or temporally misaligned mentions may still occur, for example due to entity normalization across survey papers, retrospective citations, or ambiguous naming conventions. For this reason, ACLBot is intended to support exploratory analysis and hypothesis generation rather than authoritative bibliometric reporting, and users are encouraged to inspect the underlying evidence paths returned by the system.

The quality and completeness of the knowledge graph also place inherent limits on system performance. While the current graph covers over thou-

sands of papers each of which contain many passages, extraction errors, missing metadata, and evolving terminology can affect downstream analyses. We did not conduct systematic ablation studies on KG coverage or temporal slicing, which would be required to quantify the impact of partial or noisy knowledge graphs on system behavior. Such controlled comparisons, including contrastive evaluations against LLM-only baselines, are left for future work. The evaluation presented in this paper focuses on formative user feedback rather than task-level retrieval benchmarks.

Future work will explore strategies to improve ACLBot's robustness. Potential directions include prompt optimization via automated search, few-shot prompt calibration, and integration of structured knowledge sources to reduce dependence on prompt wording. We also plan to investigate hybrid architectures that combine LLM-based generation with deterministic query execution, aiming to achieve more consistent and reliable system behavior. The use of KGs to encode dialogue, domain, and document structures (Wilcock and Jokinen, 2025) is a possible approach to integrate generative models and KGs for a uniform representation for flexible dialogue management.

8. Conclusion

This paper presented a system for literature exploration in the ACL Anthology by combining a data extraction-based structured knowledge graph which is queried via an LLM to generate answers to user questions that combine the features of both these powerful components.

The results demonstrate the potential of combining LLM-based generation with structured data to produce informative and contextually relevant responses. The study highlights challenges related to prompt sensitivity and consistency, and accuracy metrics for usability.

Future efforts will focus on improving robustness, enhancing factual grounding, and expanding the range of supported query types. We believe that addressing these limitations will move the system closer to dependable, domain-adapted conversational agents for research support.

9. Acknowledgements

This work was supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Research, Technology and Space. This paper is also based on results obtained from the project, "Research and Development Project of Enhanced Infrastructures for Post 5G Information and Communication Systems" (JPNP 20017), commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

10. Bibliographical References

2008–2025. Grobid. https://github.com/kermitt2/grobid.

Shashank Agarwal, Qian Chen, and William Yang Wang. 2023. Knowledge graph-augmented language models for question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5234–5247.

Sören Auer, Tobias Mann, Ana Marjanovic, Claus Stadler, Jennifer D'Souza, Markus Häberle, Akhilesh Palavalli, Maximilian Prinz, and Rafael Rojas. 2020. The open research knowledge graph: Towards machine actionability in scholarly communication. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 1–4.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research. ACL Anthology.

Patrik Christen, Vasileios Iosifidis, and Jens Lehmann. 2023. A conversational interface for scholarly knowledge graphs. In *Companion Proceedings of the ACM Web Conference 2023*, pages 128–132.

Emil Eifrem, Anders Östling, Johan Svensson, Tobias Ivarsson, and Peter Neubauer. 2010. Neo4j: The graph database. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD '10), Demonstration Track*. ACM.

Nigel Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*, pages 1433–1445. ACM.

Steve Harris and Andy Seaborne. 2013. Sparql 1.1 query language. Recommendation, W3C.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*. Curran Associates, Inc.

Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, and Suporn Pongnumkul. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, pages 111–119.

Neo4j Inc. 2023. Building a knowledge graph-based chatbot with gpt-3 and neo4j. https://neo4j.com/blog/developer/knowledge-graph-based-chatbot-with-gpt-3-and-neo4j/.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Transactions on Graph Data and Knowledge (TGDK)*, 1(1):2:1–2:38.

Shaurya Rohatgi. 2022. Acl anthology corpus with full text. Github.

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. 2024. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7963–7977, Miami, Florida, USA. Association for Computational Linguistics.

Phillip Schneider, Nils Rehtanz, Kristiina Jokinen, and Florian Matthes. 2023. 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.

Yinan Shen, Seung-won Hwang Lee, Jaewoo Choi, and Edward Choi. 2020. Discovering knowledge graph schema from question-answer pairs. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 868–877.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246.

Graham Wilcock and Kristiina Jokinen. 2025. Integrating conversational entities and dialogue histories with knowledge graphs and generative AI. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 290–298, Bilbao, Spain. Association for Computational Linguistics.