From Chat to Publication Management: Organizing your related work using BibSonomy & LLMs

Tom Völker
tom.voelker@informatik.uni-wuerzburg.de
Center for Artificial Intelligence and Data Science (CAIDAS), Institute for Computer Science, University of Würzburg
Würzburg, Germany

Jan Pfister
pfister@informatik.uni-wuerzburg.de
Center for Artificial Intelligence and Data Science (CAIDAS), Institute for Computer Science, University of Würzburg
Würzburg, Germany

Tobias Koopmann
koopmann@informatik.uni-wuerzburg.de
Center for Artificial Intelligence and Data Science (CAIDAS), Institute for Computer Science, University of Würzburg
Würzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
Center for Artificial Intelligence and Data Science (CAIDAS), Institute for Computer Science, University of Würzburg
Würzburg, Germany

ABSTRACT

The ever-growing corpus of scientific literature presents significant challenges for researchers with respect to discovery, management, and annotation of relevant publications. Traditional platforms like Semantic Scholar, BibSonomy, and Zotero offer tools for literature management, but largely require manual laborious and error-prone input of tags and metadata. Here, we introduce a novel retrieval augmented generation system that leverages chat-based large language models (LLMs) to streamline and enhance the process of publication management. It provides a unified chat-based interface, enabling intuitive interactions with various backends, including Semantic Scholar, BibSonomy, and the Zotero Webscraper. It supports two main use-cases: (1) Explorative Search & Retrieval - leveraging LLMs to search for and retrieve both specific and general scientific publications, while addressing the challenges of content hallucination and data obsolescence; and (2) Cataloguing & Management - aiding in the organization of personal publication libraries, in this case BibSonomy, by automating the addition of metadata and tags, while facilitating manual edits and updates. We compare our system to different LLM models in three different settings, including a user study, and we can show its advantages in different metrics.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; **Web-based interaction**; • **Applied computing** → **Document management**; Digital libraries and archives; Document metadata; • **Information systems** → **Search interfaces**; *Application servers.*

KEYWORDS

ChatGPT, Publication Management, RAG, Academic Search

ACM Reference Format:
Tom Völker, Jan Pfister, Tobias Koopmann, and Andreas Hotho. 2024. From Chat to Publication Management: Organizing your related work using BibSonomy & LLMs. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627508.3638298

1 INTRODUCTION

The constantly expanding scientific landscape makes it a significant challenge for researchers to efficiently find, manage and keep track of relevant literature. To this end, various platforms, such as Semantic Scholar [10], BibSonomy [3], or Zotero [8] offer a range of tools to search, categorize, and index these publications. Unfortunately, manual and consistent addition of metadata and tags can be error-prone and time-consuming.

Fortunately, recent advances in chat-based large language models offer a promising opportunity to support and streamline this publication management process. In this context, we introduce a retrieval augmented generation system tailored for scientific research and publication management. This system is designed to enable the user to seamlessly interact with various backends, via an intuitive chat interface. Thus, it provides a simple, cohesive, and intuitive interface, enabling the user to interact with different platforms using natural language. Currently, three backends are supported: SemanticScholar, BibSonomy, and the Zotero Webscraper. Two separate but closely related tasks are accomplished by our system: seeking out new or known scientific publications, and managing a personal virtual library of self-posted literature.

(Explorative) Search & Retrieval This use case involves searching for specific academic literature that the user wants

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638298

to retrieve or conducting a broader, more general search when the sought-after material is not yet known to the user. To accomplish this, we utilize recent advances in chat-based large language models. These pretrained models have been shown to possess vast amounts of general knowledge, enabling them to confidently describe complex concepts [4], reason over difficult problems [6], and can even provide details on scientific publications included in their training corpus (see 4). However, relying solely on these models for publication management can be problematic or flat out impossible due to numerous serious issues arising with that: First, the tendency of the models to hallucinate wrong, but plausible sounding content [2, 5, 9, 13] and second, the currency of these models, as they are only as up-to-date as the data on which they are trained. Thus, these models cannot return the most recent publication and may even invent publications that do not exist. In contrast, our system enables the model to query different knowledge bases to provide well-founded and verifiable answers.

Cataloguing & Management The second use-case centers around efficiently managing a researcher's personal publication library. This includes not only the organization and collection of publications but also the addition of correct metadata and, ideally, relevant and user-specific tags and descriptions. Although metadata can be collected from large-scale academic sources such as Semantic Scholar, manually transferring or even updating existing entries with new information, such as PDFs or a newly relevant, user-specific tag is usually tedious work. To mitigate this, our system allows for editing of posts, allowing for manual addition of metadata or automatic collection of data from other sources.

In this work, we introduce an open-source backend server[1] that enables large-language models (LLMs) to query different services. This server can be used as a retrieval engine for LLMs like ChatGPT. It allows to exploratively search for related work leveraging the open-world knowledge in LLMs, and organize self-managed publications by automating the process of adding relevant tags and metadata to publications. Additionally, we evaluate our tool in three settings, including a user study compared to other tools in this domain.

2 RELATED WORK

Large language models with transformer architectures [7, 19] represent a major advance in natural language processing, offering a multitude of potential applications [4, 6, 11]. However, a significant challenge associated with these models is their tendency to generate inaccurate or fabricated information [2, 5, 13]. To mitigate this problem and improve the reliability of these models, recent research efforts are focused on the integration of structured data. [12, 16] This aims to provide these models with structured information by providing viable facts and help the models generate accurate answers.

3 BIBSONOMY CHAT-LLM SERVICE

Our open-source backend service now builds on this concept and addresses the settings of *Search & Retrieval* as well as *Cataloguing & Management*. It enables chat-based large language models (LLMs) to communicate with different bibliography services, as depicted in Figure 1. This approach follows the "toolformer"-paradigm [16], which allows large language models to proficiently control and interact with different services. In this setting, the LLM effectively functions as a mediator between unstructured, free-form user input and structured queries to a backend system. User queries can range from simple tasks such as converting an APA-formatted citation into bibtex (Appendix A.2) to more intricate, multi-step requests like locating the Llama 2 Foundation Paper and adding it to the user's BibSonomy Account (Appendix A.3) For this demo, we deploy our service within the OpenAI-LLM-plugin ecosystem [14], but want to emphasize that any other pretrained and promptable (e.g. self-hosted) LLM can be used with our system [18].

3.1 Prompting the Model

We first describe how the model is prompted and which information is provided to the model. Technically, the entire interface is defined within a single `yaml`-file given to the model before the user interacts with it. Here, all available http-endpoints, including their respective parameters, are defined and explained to the model using natural language. The endpoints we offer are implemented in an intermediate Python server, which serves as a middle layer between the different backends and the chat model (Figure 1). After receiving the response from our service, the LLM will be prompted with this information to generate a more concise response for the user.

Furthermore, an initial system prompt is given to make the model aware of its environment (Appendix A.8). In addition to providing a brief overview of the system's functionality and target use cases, this prompt primarily serves to introduce the model to the desired mode of operation. On the one hand, the prompt explicitly explains how good search queries should be formulated, but it also points out the relevance of data integrity in the context of scientific work (to minimize hallucinations). Furthermore, it explains the desired procedure when posting content to BibSonomy (e.g. how tags and descriptions should be used, etc.). Finally, we have had great success including additional system prompts within the API responses. This approach adds appropriate suggested actions or hints to the model based on the API endpoint used (and other parameters) to make the model behave as desired.

3.2 Endpoints Provided by the Backend

(Explorative) Search & Retrieval The first use-case revolves around exploring established or undiscovered scientific work, for which we offer two simple endpoints to the model. They allow querying different backends and using the results to generate an answer, inspired by the "Retrieval Augmented Generation"-paradigm [12]. The two provided http-endpoints are called `/search` and `/details`, which can be used with a multitude of optional query parameters. They provide a unified and simple access to an easily extendable list of backends

[1] https://bitbucket.org/bibsonomy/bibsonomy-llm-scientific-retrieval-plugin

(currently: BibSonomy, SemanticScholar, and Zotero scrapers). Support can be easily extended to other back-ends, given they provide API endpoints conceptually matching either of these endpoints. *The "/search"-endpoint* is designed to handle a variety of query inputs, accepting free-form text queries or parameters such as title, authors, or keywords. Upon receiving a request, the service transforms it internally into native query formats compatible with various supported backends, specifically for this endpoint, BibSonomy and SemanticScholar. To mimic a researcher's approach, the model can also enrich the search by including supplementary search terms closely related to the main query, allowing for a more holistic understanding of the topic. By default, both backends are queried for the top results for each query, with the results then getting merged and re-ranked within our backend, before they are returned to the chat model. We concatenate the results from both backends and reorder them according to the results BM25 scores [15]. Here, the system considers the query type and the number of platforms on which a result appears. In particular, results sourced from the primary query are given substantially more weight than those from supplementary queries, ensuring the relevance of returned results. The model can also choose to explicitly query a single backend or search for a different number of results by specifying the query parameters. Given the limited context length of LLMs [19], they cannot handle an extremely long input. Extensive data, such as numerous relevant publications, returned by the backend may lead to struggles with the response length. We thus provide two parameters, which are by default set to be rather strict and have to be explicitly raised by the model to prevent it from being quickly overwhelmed by long results: (1) the number of results returned, and (2) the granularity of the returned information, which is "basic" by default (title, author, year, number of citations) but can be specified to be more "verbose", containing more metadata information provided by the respective platform. Furthermore, a unique platform-specific identifier is returned for each publication, providing a short handle to the model to further reference the specific publication when communicating with our backend. This has two benefits: (1) makes it easier for the model to clearly reference a result with only comparatively few tokens added to the context window, (2) while also speeding up the interaction as less tokens have to be generated when formulating the requests that are sent to either endpoint. For the *"/details"-endpoint* this unique identifier is enough to retrieve all metadata information provided by the platform. Furthermore, it can also resolve various other commonly used identifiers such as DOIs, arXiv-IDs, ACL-IDs, etc. , or even arbitrary URLs, which are then resolved using the well-maintained Zotero WebScraping server [21]. This enables unique workflows, like providing a URL to a publication, for which the model then fetches basic metadata using Zotero, looks up the correct bibtex using Semantic Scholar, and posts it with all gathered metadata to the users' personal BibSonomy library (Appendix A.4).

Publication Management Once a publication is in the user's personal BibSonomy library, it can be further managed either "traditionally" via the BibSonomy interface, or again via this system using the chat interface. This expands the model's capabilities to go beyond a simple "read-only" mode: Here *the "/search"-endpoint* provides the ability to search within the users' private library, either filtered by keywords, or manually assigned keyword tags. Furthermore, it is possible to edit posts, for example, to add or update bibliography metadata, change user-given tags, or even upload an associated PDF file. Since our backend offers access to the user's previously assigned tags, the LLM is capable of comprehending the user-specific tagging system that has already been employed. Consequently, it can accurately apply tags to existing or newly added literature in a manner consistent with the user's approach, utilizing pre-existing tags or introducing new ones as needed and appropriate. Furthermore, when adding a new publication, the model can add a short description to the publication, briefly describing the content or relevant context of the publication. Whenever a publication is added or modified using a language model, it is automatically assigned the tags `posted_with_chatgpt` or `edited_with_chatgpt`, thus the users are aware that LLMs have been used in the processing of this entry.

4 EVALUATION

Our system has been successfully installed and used by more than 350 unique users. Now, we will evaluate our chat integration by comparing it qualitatively and quantitatively with similar tools, namely (1) Vanilla ChatGPT-4 (2) You.com [20] (3) ChatGPT-4 with Bing (4) ChatGPT-4 with ScholarAI [17] (5) and finally ChatGPT-4 with BibSonomy (ours). While some systems, like ScholarAI and ours, cater to academic research, others, such as Bing and You.com, rely on expensive internet browsing. Vanilla ChatGPT-4 relies solely on its training data and is used as a baseline. ScholarAI, sharing a similar concept with our plugin, is our main point of comparison. However, it focuses more on detailed data for specific papers than on an overall topic perspective. Moreover, their publication management options only extend to a rudimentary Zotero integration, requiring manual API key inputs. Additionally, ScholarAI is not open source, imposing paywalls and limits on searches, and lacks transparency on its processes and data. We evaluate our backend's search and retrieval capabilities under three different aspects: user opinion, determinism of query results, and inference time.

User Studies: We prompt each of the five systems introduced above with the same seven queries (Appendix A.7). and ask 17 machine learning PhD candidates to rank the results against each other. The prompts were carefully generated and selected by us beforehand to cover different levels of complexity and research approaches and focus on retrieving specific articles based on topics and authors (Appendix A.7). Overall 219 votes were cast between the replies of two random models to the identical prompt[2]. The decision is a blind (i.e. the experts do

[2]The website used, as well as the code for the evaluation, as well as all model prompts and replies can be found in the repository.

not know which LLM created which answer), pairwise comparison polling three different aspects: Which response fits **intuitively better** for the query, is more **scientifically valid** (i.e. no hallucinations) and which response is more **up to date**. From Table 1 we find that the answers provided by our system are clearly favored over all other responses (favored in about 90% of the pairwise comparisons) in the three aspects evaluated. ChatGPT4 with ScholarAI, as well as Bing, is closest to our system, as it is preferred in 53% to 59% of comparisons regarding how "intuitive" the result matches the user prompt and the "validity" of the reply. Furthermore, the former is preferred in 73% of the cases for "recency", while the latter is preferred only in 55% of the comparisons here. The other systems are generally preferred in less than 40% of pairwise comparisons. Note that the metrics exceed 100% due to the cumulative effect of pairwise comparisons across the five models, resulting in an aggregated total that surpasses the conventional percentage threshold.

The queries used can be divided into two categories: topic-related prompts and author-related prompts. For topic-related prompts, models with unlimited internet access, such as the Bing model, demonstrate a strong performance. Their strength lies in the ability to search the Internet, accessing diverse sources such as blogs and other non-traditional academic platforms. Yet, in direct comparisons, our system still emerges as the preferred choice, highlighting its proficiency in focused topic searches.

When considering author-specific prompts, the landscape changes. Here, the breadth of information accessed by these internet-dependent models can sometimes work against them, leading to less coherent outputs when handling expansive data requests. While Scholar AI, as another database-reliant system, presents commendable results, our plugin consistently emerges as the top choice. This highlights the benefit of using a backend with structural data access, which emphasizes the benefit of our system approach.

Determinism of Query Results Secondly, we evaluated reply determinism of different systems across several executions with an identical prompt. We argue that this is a core requirement of a reliable retrieval system, as it should not return a different random subset of relevant publications every time the user queries. For this assessment, we compare the four systems with access to external resources in two settings. We prompt the models to retrieve three recent and three highly influential publications of a specific author (Appendix A.7.5) , as well as identifying the Palm2 [1] publication, as "a new model by one of the original BERT authors" (Appendix A.7.2) We evaluated the number of identical linked publications for the same query in five iterations . Here, the BibSonomy plugin again consistently returns the correct publication in every run, showcasing the ability to access structured data and a coordinated search strategy. Although ScholarAI's model is capable of consistently citing the most influential items, it only once provided the desired combination of both recent and influential publications and failed to recognize the Palm2 publication once. On the contrary, GPT-4 with Bing and You.com showed general inconsistencies in their responses. Even when browsing similar sources during its search, the final responses vary widely, such that instead of returning six publications as requested, the model frequently returns CVs and other unrelated information, often omitting the desired publications. Both never correctly identified the Palm2 publication and often erroneously returned nonrelated work. You.com was unable to provide consistent responses and additionally regularly returns non-existent publications and other hallucinations.

Inference Time Finally, we measure the average time-to-response of the different engines for the two prompts mentioned above. Here, the time from the "send" action was measured until the model started to respond. Vanilla GPT-4 and You.com were excluded based on their subpar response quality, as they were preferred in less than 50% of the user evaluation instances. While the BibSonomy and ScholarAI systems yield results within a similar time frame, GPT-4 with Bing requires on average more than three times longer to retrieve its data, occasionally exceeding two minutes before answering. We argue that this significantly affects the usability of the model and shows the need for a use-case-tailored search backend.

5 CONCLUSION

We present the BibSonomy Chat-LLM Service, which streamlines and simplifies publication management for researchers. The plugin accesses different back-ends to conduct smart searches for new, related work from large data sources like Semantic Scholar. Furthermore, it seamlessly integrates with the BibSonomy publication management system, enabling it to add, edit, and fetch data from each user's personal collection. By using BibSonomy's tagging system, the plugin can intelligently employ preexisting tags or incorporate new ones suggested by users during chat interactions. This results in a more streamlined search and management process for scientific publications. In a user evaluation, we find that our system provides responses that are not only intuitively superior, but also consist of more recent publications and does not fabricate nonexistent publications. Additional steps could be to integrate additional data sources, ideally with publicly available API access. Furthermore, with an increased input context size, it could become feasible to post entire paragraphs of publications to extract all citations and add them to BibSonomy. Lastly, given a tagged set of publications in a BibSonomy collection, the LLMs could generate a summary of these works and put them into perspective.

ACKNOWLEDGMENTS

This work is supported by the MOTIV research project funded by the Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities. The authors are responsible for the content of this publication.

REFERENCES

[1] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey,

Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. PaLM 2 Technical Report. *CoRR* abs/2305.10403 (2023). https://doi.org/10.48550/ARXIV.2305.10403 arXiv:2305.10403

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. 2010. The social bookmark and publication management system bibsonomy. *The VLDB Journal* 19, 6 (01 Dec 2010), 849–875. https://doi.org/10.1007/s00778-010-0208-4

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *CoRR* abs/2107.03374 (2021). arXiv:2107.03374 https://arxiv.org/abs/2107.03374

[6] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 3882–3890. https://doi.org/10.24963/IJCAI.2020/537

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Peter Fernandez. 2011. Zotero: information management software 2.0. *Library Hi Tech News* 28, 4 (01 Jan 2011), 5–7. https://doi.org/10.1108/07419051111154758

[9] Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. arXiv:2303.16104 [cs.CL]

[10] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul L Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *ArXiv* abs/2301.10140 (2023). https://api.semanticscholar.org/CorpusID:256194545

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[12] 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 *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[13] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3214–3252. https://doi.org/10.18653/V1/2022.ACL-LONG.229

[14] OpenAI. 2023. ChatGPT plugins. https://openai.com/blog/chatgpt-plugins Accessed: 2023-10-30.

[15] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[16] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761 [cs.CL]

[17] ScholarAI. 2023. ScholarAI. https://scholarai.io Accessed: 2023-10-30.

[18] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. 2023. OpenAgents: An Open Platform for Language Agents in the Wild. arXiv:2310.10634 [cs.CL]

[19] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]

[20] You.com. 2023. You.com. https://you.com Accessed: 2023-10-30.

[21] Zotero. 2023. Zotero Translation Server. https://github.com/zotero/translation-server Accessed: 2023-10-30.