PSBench: A Benchmark for Automated Academic Paper Search

Junyi Hou[a], Haodong Zhao[b], Bingsheng He[a]

[a] *National University of Singapore* junyi.h@comp.nus.edu.sg, hebs@comp.nus.edu.sg

[b] *Shanghai Jiao Tong University* zhaohaodong@sjtu.edu.cn

* Presenting author

1. Abstract

The exponential growth of academic publications has created information overload [1], making it increasingly difficult for researchers to locate relevant literature. Although Google Scholar remains the predominant academic search engine [2], its keyword-based retrieval system shows limitations in fast-moving domains. A key challenge lies in mapping researchers' complex information needs to search queries, as a simple keyword combination rarely captures the full scope of modern research. As highlighted in recent studies [1], there is an urgent need to improve how we search and the tools we use to improve discovery capabilities.

Advances in Large Language Models (LLMs) present opportunities for automated scientific discovery. Several attempts have been made to use LLMs as autonomous agents to search and retrieve academic works [3, 4]. However, current benchmarks [3, 5] typically focus on natural language queries, while actual literature searches are usually conducted with keyword-based queries, iterative refinement, and domain-specific heuristics [5]. As a result, it remains uncertain whether LLM-based retrieval systems truly enhance paper discovery for researchers in practice.

To address this gap, we introduce a novel benchmark that systematically evaluates literature search tools. Using papers from ICLR 2024 [6], a leading conference in AI research, we generate ground truth retrieval lists from the references of each paper. This approach allows us to objectively assess the retrieval accuracy of different search engines.

Additionally, we introduce a new retrieval method that combines semantic search with domain-specific knowledge graphs. Our method captures both text meaning and structural connections between papers, delivering more relevant results. Testing shows it surpasses standard methods in precision and recall.

In summary, our contributions include:

- A novel benchmark for comprehensive evaluation of finding relevant papers on different academic literature search tools.
- An innovative retrieval method combining semantic search with domain-specific knowledge graphs.
- Empirical evidence of improved retrieval accuracy over baseline methods.

2. Impacts

Beyond our methodological contributions, our work offers clear benefits for researchers at all stages. For new scholars, improved academic search tools simplify literature searches, helping them quickly gather the information to write stronger manuscripts and increase their chances of paper acceptance. For seasoned researchers, our context-aware retrieval systems offer accurate access to most-relevant literature—even in unfamiliar areas—making it easier to assess the novelty and significance of the latest research.

3. Related Work

Current literature search datasets include the Semantic Scholar Open Research Corpus (S2ORC) [7], which provides citation lists for a large number of academic papers. However, the quality of individual papers and their reference lists can vary, which may not guarantee the relevance and accuracy needed for a benchmark.

BigSurvey [8] and Surfer100 [9] are datasets focused on summarizing academic papers, which is a different task from literature search. While they provide valuable resources for understanding paper content, they do not directly contribute to the evaluation of search tools.

Recent works such as AutoScholarQuery [3] and LitSearch [10] have introduced datasets for scientific literature search. AutoScholarQuery employs instruction-like natural language queries for paper search, which may not fully reflect real human search behavior often characterized by keyword-based searches. In contrast, LitSearch offers a retrieval benchmark, constructed using questions generated by GPT-4 [11] from citation contexts and manually written by authors, focusing on finding papers that answer specific research questions. Our benchmark, however, is designed to mimic real-world research queries by utilizing keyword-based searches and assesses the retrieval of relevant papers based on keywords, paper titles, or abstracts.

In summary, existing datasets and benchmarks have limitations in terms of reflecting real human search behavior and ensuring the quality and relevance. Our proposed benchmark addresses these gaps by utilizing high-quality papers from ICLR 2024 and their reference lists to create a robust evaluation framework for literature search tools.

Acknowledgments

References

[1] Michael Gusenbauer and Neal R. Haddaway. What every researcher should know about searching – clarified concepts, search advice, and an agenda to improve finding in academia. *Research Synthesis Methods*, 12(2):136–147, 2021.

[2] David Nicholas, Cherifa Boukacem-Zeghmouri, Blanca Rodríguez-Bravo, Jie Xu, Anthony Watkinson, A. Abrizah, Eti Herman, and Marzena Świgoń. Where and how early career researchers find scholarly information. *Learned Publishing*, 30(1):19–29, 2017.

[3] Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. Pasa: An llm agent for comprehensive academic paper search, 2025.

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Hao Kang and Chenyan Xiong. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents, 2025.

[6] *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. 2024.

[7] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[8] Shuaiqi LIU, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. Generating a structured summary of numerous academic papers: Dataset and method. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-2022, page 4259–4265. International Joint Conferences on Artificial Intelligence Organization, July 2022.

[9] Irene Li, Alex Fabbri, Rina Kawamura, Yixin Liu, Xiangru Tang, Jaesung Tae, Chang Shen, Sally Ma, Tomoe Mizutani, and Dragomir Radev. Surfer100: Generating surveys from web resources, Wikipedia-style. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5388–5392, Marseille, France, June 2022. European Language Resources Association.

[10] Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[11] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski,

Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.