A Pipeline to Bootstrap the Evaluation of Retrieval-Augmented Generation for the Automation of Systematic Reviews in Computer Science

Pierre Achkar
Leipzig University; Fraunhofer ISI

Tim Gollub
Bauhaus-Universität Weimar

Arno Simons
TU Berlin

Harrisen Scells
University of Tübingen

Maik Fröbe
Friedrich-Schiller-Universität Jena

Martin Potthast
Kassel University; hessian.AI; ScaDS.AI

Abstract

Automating systematic reviews (SRs), i.e., evidence-driven analyses under explicit protocol constraints, is a natural target for retrieval-augmented generation and deep research agents, yet existing benchmarks evaluate isolated subtasks or assume fixed evidence inputs. We introduce RAG4SR-CS-200, a benchmark of 200 computer science systematic reviews designed for protocol-driven systematic review automation. Each instance comprises review objectives, research questions, eligibility criteria, cleaned full-text review structure, references, and extracted tables. These elements support evaluation across key tasks in systematic review creation such as literature retrieval, eligibility screening, citation-grounded review generation, and structured table generation, in both stage-wise and end-to-end settings. RAG4SR-CS-200 provides a foundation for developing more reliable and diagnosable deep research agents for scientific evidence synthesis. Code and data are publicly available[1].

1 Introduction

Systematic reviews are the gold standard for scientific evidence synthesis (Mulrow, 1994; Gough et al., 2012), given their protocol-driven, transparent, and reproducible methodology. Yet producing a high-quality systematic review is expensive in both time and effort (Borah et al., 2017). The process also follows clear stages: define the objective and research questions, retrieve candidate studies, screen studies for eligibility, and synthesise accepted evidence into a structured, cited report (Lefebvre et al., 2019; Liberati et al., 2009). This staged process aligns well with multi-stage Retrieval-Augmented Generation (RAG) pipelines (Lewis et al., 2020) and deep research agents (Zhang et al., 2025), where intermediate outputs can be evaluated at each step.

Recent work has made strong progress in SR automation, including end-to-end systems such as otto-SR (Cao et al., 2025) and multi-agent SLR pipelines (Sami et al., 2024). However, existing benchmarks still focus mainly on isolated subtasks. For retrieval and screening, resources such as CLEF TAR (Kanoulas et al., 2017, 2018, 2019), the SysRev Query Collection (Scells et al., 2017), and CSMeD (Kusa et al., 2023) have enabled advances in query formulation and citation screening. For synthesis, datasets such as SciReviewGen, SurveySum, and SumSurvey (Kasanishi et al., 2023; Fernandes et al., 2024; Liu et al., 2024), together with frameworks such as AutoSurvey and SurveyForge (Wang et al., 2024; Yan et al., 2025), mainly target narrative reviews (surveys). Unlike systematic reviews, narrative reviews are generally expert-curated and more subjective, and they do not require protocol-defined eligibility criteria, explicit research questions, or reproducible study selection. Related work also studies structured outputs, such as hierarchical catalogue generation (Zhu et al., 2023) and survey table generation (Chen et al., 2024). Still, some benchmarks assume fixed or pre-selected evidence, leaving protocol-driven retrieval and eligibility decisions outside scope and limiting end-to-end error diagnosis across retrieval, screening, generation, and structured outputs.

To address this gap, we introduce **RAG4SR-CS-200**, a benchmark of 200 computer science systematic reviews for evaluating RAG and deep research agent pipelines for systematic review automation. Each instance combines structured inputs (topic, research objective, research questions, eligibility criteria, and temporal restrictions) with the full re-

[1] https://github.com/webis-de/rag4sr-cs-200

view text, in-text citation markers, extracted tables, and a resolved reference list. RAG4SR-CS-200 supports evaluation across four stages: **(i)** literature retrieval, **(ii)** eligibility screening, **(iii)** citation-grounded review generation, and **(iv)** structured table generation. The design supports both stage-wise and end-to-end evaluation under realistic retrieval conditions. The benchmark specifies its format, construction pipeline, and evaluation protocol. As the first release in the RAG4SR-X series, designed to expand to additional domains, it provides a foundation for cross-domain benchmarking.

2 Related Work

Prior work on systematic review automation has largely focused on individual pipeline stages. For retrieval and screening, widely used resources such as CLEF TAR (Kanoulas et al., 2017, 2018, 2019), the SysRev Query Collection (Scells et al., 2017), Seed Studies (Wang et al., 2022), SWIFT-Review (Howard et al., 2016), SR Updates (Alharbi and Stevenson, 2019), and AutoBool (Wang et al., 2025a) have enabled progress in Boolean query formulation and citation screening (Scells et al., 2021; Wang et al., 2023, 2025b; Cohen et al., 2006; Wallace et al., 2010; Miwa et al., 2014). However, these datasets mainly evaluate early-stage tasks rather than the full end-to-end review pipeline.

At the synthesis stage, several resources target downstream generation quality. SciReviewGen (Kasanishi et al., 2023) and SurveySum (Fernandes et al., 2024) generate survey sections from cited papers, while SumSurvey (Liu et al., 2024) targets long-document summarisation of full survey papers. SurveyForge (Yan et al., 2025) is a pipeline for producing full survey drafts with structure-aware generation and memory-guided retrieval. SurveyBench (Sun et al., 2025) and SGSimEval (Guo et al., 2025) are evaluation resources that score generated surveys beyond lexical overlap, including structure, content quality, and reference quality. HiCatGLR (Zhu et al., 2023) and Survey Table Generation (Chen et al., 2024) target structured outputs, namely hierarchical outlines and comparison tables. These resources are valuable, but many rely on fixed inputs, pre-collected corpora, or pre-selected references; protocol-driven retrieval and eligibility screening are typically outside the evaluated scope. RAG4SR-CS-200 addresses this gap by targeting protocol-driven systematic review automation, with unified evaluation across retrieval, screening, citation-grounded generation, and structured outputs.

Table 1 summarises representative resources by benchmark coverage and highlights the remaining gap: unified support for stage-wise and end-to-end evaluation under a common protocol.

3 Dataset Construction

This section describes how we construct RAG4SR-CS-200 from an existing large-scale systematic review resource and transform it into benchmark-ready artefacts for SR automation.

3.1 Goal and Scope

RAG4SR-CS-200 is designed for protocol-driven *systematic review* (SR) automation. The benchmark supports unified evaluation across retrieval, eligibility screening, citation-grounded generation, and table generation. In contrast to survey-focused resources, we curate reviews exposing SR-specific method signals, including a stated objective, research questions, and eligibility criteria.

3.2 Collection and Selection Pipeline

The construction starts from a large-scale dataset of systematic reviews that we collected from OpenAlex (Priem et al., 2022); this upstream resource is part of our ongoing unpublished work. OpenAlex indexes a wide range of scholarly documents across domains and provides broad citation metadata coverage (Culbert et al., 2025). On top of this collection, we perform structured extraction over reviews with accessible full text to identify explicitly reported review-protocol artefacts. From this source pool, we derive a computer-science-focused subset by applying strict SR-oriented inclusion filters. Specifically, a review is retained only if it provides: *(i)* an explicit review objective, *(ii)* explicit research questions, and *(iii)* explicit eligibility criteria. We exclude records missing any of these protocol artifacts, outside computer science scope, or lacking usable full text after parsing. The resulting subset contains 200 computer science systematic reviews.

3.3 Data Extraction and Normalisation

For the selected reviews, parsed markdown is already available from the upstream pipeline (generated from PDF using *PaddleOCR-VL*,[2]). However, this parsed text remains noisy for benchmark use,

[2] https://github.com/PaddlePaddle/PaddleOCR

Resource	Domain	R	S	G	T	E2E
CLEF TAR (Kanoulas et al., 2017, 2018, 2019)	Biomedical	X	X			
SysRev Query Collection (Scells et al., 2017)	Biomedical	X				
Seed Studies (Wang et al., 2022)	Biomedical	X				
SWIFT-Review (Howard et al., 2016)	Biomedical		X			
SR Updates (Alharbi and Stevenson, 2019)	Biomedical		X			
AutoBool (Wang et al., 2025a)	Biomedical	X				
SciReviewGen (Kasanishi et al., 2023)	CS/NLP			X		
SurveySum (Fernandes et al., 2024)	CS/NLP			X		
SurveyForge (Yan et al., 2025)	CS/NLP	X		X		
SurveyBench (Sun et al., 2025)	CS/NLP			X		
SGSimEval (Guo et al., 2025)	CS/NLP			X		
HiCatGLR (Zhu et al., 2023)	CS/NLP			X		
Survey Table Generation (Chen et al., 2024)	CS/NLP				X	
RAG4SR-CS-200 (ours)	CS	X	X	X	X	X

Table 1: Representative related work by domain and benchmark coverage. R/S/G/T indicate whether a resource explicitly evaluates retrieval, eligibility screening, citation-grounded generation, or structured table generation. E2E indicates joint evaluation of the full protocol under one benchmark

as headings, citations, and tables are not consistently normalised across reviews. For RAG4SR-CS-200, we therefore apply an additional LLM-assisted cleanup workflow. We first run *DeepSeek-V3.1*[3] for structural cleanup and citation normalisation, then manually compare each cleaned markdown file against the source PDF. During this stage, we enforce consistent heading structure, remove conversion artefacts (e.g., extra whitespace, incorrect line breaks), and normalise in-text citations into numeric markers to make citation grounding comparable across reviews.

After text cleanup, we parse review reference lists with *Anystyle*[4] to extract bibliographic fields, including title, author list, DOI, and publication year. We then normalise reference metadata against *Crossref* to ensure consistent formatting and to obtain OpenAlex identifiers for alignment with the indexed reference corpus. This produces a cleaned, linkable reference layer for retrieval and grounding evaluation.

3.4 Benchmark Schema and Quality Control

Each review is distributed as a structured JSON record, an accompanying markdown file containing extracted tables and the reference list with OpenAlex identifiers. At the JSON level, each instance includes: review identifier, high-level metadata (e.g., section and table counts), section/subsection hierarchy with cleaned text, normalised in-text citation markers, table placeholders in context, and a table index that links placeholders to source locations. The table artefact stores full markdown-renderable table content. Figure 1 shows a compact per-review JSON example.

```
{
"id": "W1505282872",
"metadata": { "title": ..., "n_sections": 3,
"n_subsections": 3, "n_tables": 19 },
"sections": [
{
"section_id": "s_1",
"section_label": "1. Introduction",
"text": ...,
"citations": ["W1993563915", "W2123444177",
...],
"tables_in_text": [ ... ],
"subsections": [ ... ]
}
],
"tables": [
{ "table_id": "tbl_01", "placeholder":
"{{TABLE:tbl_01}}",
"source_section_id": "s_2",
"source_subsection_id": "s_2_1" }
],
"tables_file": "W1505282872_tables.md"
}
```

Figure 1: Compact example of the per-review JSON schema used in RAG4SR-CS-200, based on the instance in `example_data/W1505282872.json`.

Quality control combines automatic and manual checks. Automatic checks validate schema consistency, citation-marker formatting, and section-table linkage integrity. Manual checks compare cleaned markdown with the original PDF to ensure that section boundaries, citation placement, and table references remain faithful to the source document.

3.5 Statistics and Benchmark Coverage

RAG4SR-CS-200 contains 200 computer science systematic reviews with harmonised structure and citation annotations. Each instance supports all

[3] https://api-docs.deepseek.com/news/news250821

[4] https://anystyle.io/

benchmark stages: retrieval (through linked reference metadata), screening-oriented protocol artefacts (objective, research questions, eligibility criteria), citation-grounded generation (through normalised in-text citations and reference alignment), and table generation (through extracted markdown tables). On average, each review contains 77 references, 7 sections, 11 subsections, and 7 tables. In total, the corpus contains 15,498 references, of which 12,871 are matched to OpenAlex metadata, corresponding to 83.1% alignment coverage.

4 Discussion

This section summarises the benchmark scope and discusses why it is useful and challenging for RAG and agentic deep research systems.

4.1 Benchmark Scope

RAG4SR-CS-200 is intended for evaluating RAG and agentic pipelines for protocol-driven systematic review automation. The benchmark supports controlled comparisons at both individual stages and full end-to-end settings across the four tasks. At the retrieval stage, the references aligned with OpenAlex identifiers make it possible to assess whether a system can identify candidate studies relevant to the review objective and research questions. Eligibility screening is supported by protocol-defining inputs such as the review objective, research questions, and eligibility criteria, making it possible to assess whether studies are correctly retained or excluded under explicit review constraints. Citation-grounded generation can be assessed using the cleaned full-text review structure and normalised citation markers, which allow testing of whether generated synthesis remains grounded in identifiable evidence. The extracted markdown tables provide a target for assessing whether a system can produce faithful tabular summaries from the underlying review content. Taken together, these components make RAG4SR-CS-200 suitable for studying deep-research-style agents that must coordinate retrieval, reasoning, grounding, and structured synthesis over long scientific documents.

4.2 Challenges and Evaluation

Systematic review automation is a demanding setting for RAG systems and deep research agents. Unlike short-form question answering, the task requires sustained multi-step reasoning under explicit protocol constraints: a system must identify relevant evidence, respect inclusion criteria, preserve traceability to source documents, and produce outputs whose claims can be inspected against cited references. Errors at early stages, such as weak retrieval or incorrect screening, propagate into later synthesis and can silently degrade the final review. Stage-wise analysis is therefore important for diagnosing where failures begin and how they affect downstream outputs.

RAG4SR-CS-200 is designed to expose exactly these failure modes. By combining protocol signals, cleaned full-text structure, normalised citations, and tables, the benchmark supports analysis of whether a method retrieves the right evidence, grounds claims correctly, and produces faithful structured outputs. This supports both component-level analysis and stress-testing of end-to-end agentic systems for scientific research workflows. To support concrete task use, we release structured per-review JSON files, extracted table files, aligned reference metadata, and pipeline scripts in the public repository. Evaluation can be conducted at both component and end-to-end levels. At the component level, retrieval can be measured with standard retrieval metrics (e.g., precision and recall), screening can be evaluated against the references included in the review, generation can be assessed with lexical or semantic similarity measures, citation quality can be assessed with citation-specific metrics, and table generation can be evaluated at cell-, row-, and table-level granularity using lexical or semantic overlap with reference tables. End-to-end evaluation can then assess faithfulness, completeness, and traceability of the final synthesis.

5 Conclusion

We introduced RAG4SR-CS-200, a benchmark of 200 computer science systematic reviews for evaluating protocol-driven systematic review automation with RAG and agentic systems. The benchmark supports unified evaluation across retrieval, screening, citation-grounded generation, and structured table generation, enabling stage-wise and end-to-end analysis under realistic review constraints. As the first release in the planned RAG4SR-X series, it provides a foundation for broader cross-domain benchmarks and more reliable deep research agents for scientific evidence synthesis.

6 Limitations

The current release is restricted to computer science systematic reviews, so transfer to other scientific domains should not be assumed without further validation. The benchmark also inherits residual noise from PDF parsing, LLM-assisted cleanup, reference parsing, and citation resolution despite manual checks. Because this release prioritises data curation and normalisation rather than subjective annotation, we report procedural quality control instead of inter-annotator agreement. The release provides OpenAlex identifiers for references included in each review, but does not redistribute the corresponding abstracts or full-text documents; users must source this reference content when building retrieval corpora or running end-to-end experiments beyond the released artefacts.

References

Amal Alharbi and Mark Stevenson. 2019. A dataset of systematic review updates. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1257–1260. ACM.

Rohit Borah, Andrew W. Brown, Patrice L. Capers, and Kathryn A. Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ Open*, 7.

Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, Matthew Cecere, Anabel Selemon, Jason Sang, Ling Xi Gong, Robert Kloosterman, Scott Jiang, Richard Saleh, Denis Margalik, James Lin, Jane Jomy, Jerry Xie, David Chen, Jaswanth Gorla, Sylvia Lee, and 14 others. 2025. Automation of systematic reviews with large language models. *medRxiv*.

Po-Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Survey table generation from academic articles. In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 190–197.

Aaron M. Cohen, William R. Hersh, K. Peterson, and Po-Yin Yen. 2006. Research paper: Reducing workload in systematic review preparation using automated citation classification. *J. Am. Medical Informatics Assoc.*, 13(2):206–219.

Jack H. Culbert, Anne Hobert, Najko Jahn, Nick Haupka, Marion Schmidt, Paul Donner, and Philipp Mayr. 2025. Reference coverage analysis of openalex compared to web of science and scopus. *Scientometrics*, 130:2475–2492.

Leandro Car'isio Fernandes, Gustavo Bartz Guedes, Thiago Laitz, Thales Sales Almeida, Rodrigo Nogueira, R.A. Lotufo, and Jayr Pereira. 2024. Surveysum: A dataset for summarizing multiple scientific articles into a survey section. In *Brazilian Conference on Intelligent Systems*.

David Gough, Sandy Oliver, and James Thomas, editors. 2012. *An Introduction to Systematic Reviews*. Sage Publications Ltd.

Beichen Guo, Zhiyuan Wen, Yu Yang, Peng Gao, Ruosong Yang, and Jiaxing Shen. 2025. Sgsimeval: A comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems. In *International Conference on Advanced Data Mining and Applications*.

Brian E. Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R. Shah, Stephanie Holmgren, Katherine E. Pelch, Vickie Walker, Andrew A. Rooney, Malcolm Macleod, Ruchir R. Shah, and Kristina Thayer. 2016. Swift-review: a text-mining workbench for systematic review. *Systematic Reviews*, 5:87.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*, volume 1866 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Evangelos Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. SciReviewGen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6695–6715, Toronto, Canada. Association for Computational Linguistics.

Wojciech Kusa, Óscar E. Mendoza, Matthias Samwald, Petr Knoth, and Allan Hanbury. 2023. Csmed: Bridging the dataset gap in automated citation screening for systematic literature reviews. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Carol Lefebvre, Julie Glanville, Simon Briscoe, Anne Littlewood, Chris Marshall, Maria-Inti Metzendorf, Anna Noel-Storr, Tamara Rader, Farhad Shokraneh, James Thomas, and L. Susan Wieland. 2019. *Searching for and selecting studies*. John Wiley & Sons, Ltd.

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*.

A. Liberati, D. Altman, J. Tetzlaff, C. Mulrow, P. Gøtzsche, J. Ioannidis, Mike Clarke, Mike Clarke, P. Devereaux, J. Kleijnen, and D. Moher. 2009. The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med.*

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651, Bangkok, Thailand. Association for Computational Linguistics.

Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, pages 242–253.

C. D. Mulrow. 1994. Systematic reviews: Rationale for systematic reviews. *BMJ*, 309(6954):597–599.

Jason Priem, Heather A. Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*, abs/2205.01833.

Malik Abdul Sami, Zeeshan Rasheed, Kai-Kristian Kemell, Muhammad Waseem, Terhi Kilamo, Mika Saari, Kari Systä, Anh Nguyen Duc, and Pekka Abrahamsson. 2024. System for systematic literature review using multiple ai agents: Concept and an empirical evaluation a preprint.

Harrisen Scells, Guido Zuccon, and Bevan Koopman. 2021. A comparison of automatic boolean query formulation for systematic reviews. *Information Retrieval Journal*, pages 3–28.

Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. 2017. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In *Proc. of SIGIR 2017*. ACM.

Zhaojun Sun, Xuzhou Zhu, Xuanhe Zhou, Xin Tong, Shuo Wang, Jie Fu, Guoliang Li, Zhiyuan Liu, and Fan Wu. 2025. Surveybench: Can llm(-agents) write academic surveys that align with reader needs? *ArXiv*, abs/2510.03120.

Byron C. Wallace, Thomas A. Trikalinos, Joseph Lau, Carla E. Brodley, and Christopher H. Schmid. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinform.*, 11:55.

Shuai Wang, Harrisen Scells, Justin Clark, Bevan Koopman, and Guido Zuccon. 2022. From little things big things grow: A collection with seed studies for medical systematic review literature search. In *SIGIR*, pages 3176–3186.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can chatgpt write a good boolean query for systematic review literature search? In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1426–1436. ACM.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2025a. AutoBool: An Reinforcement-Learning trained LLM for Effective Automated Boolean Query Generation for Systematic Reviews. *CoRR*, abs/2602.00005.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2025b. Reassessing large language model boolean query generation for systematic reviews. In *SIGIR*, pages 3296–3305.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: large language models can automatically write surveys. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. 2025. SURVEYFORGE : On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12444–12465, Vienna, Austria. Association for Computational Linguistics.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025. Deep research: A survey of autonomous research agents. *ArXiv*, abs/2508.12752.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6790–6804, Singapore. Association for Computational Linguistics.