Automating Cognitive Distillation for Expert-Level Scientific Literature Synthesis

Jingyi Chai [1 2] **Xiangrui Liu** [1 2] **Yanbo Han** [2] **Yuwen Du** [1 2] **Tian Jin** [1 2] **Yuzhi Zhang** [2] **Linfeng Zhang** [2] **Siheng Chen** [1 2]

Abstract

Scientific progress depends on synthesizing a body of literature that is now growing faster than experts can read it. Recent large-language-model systems can write survey-like text in hours, but the result still falls far short of an expert-written review. Here we present SurveyMaster, an artificial intelligence system that generates expert-level literature reviews across scientific disciplines. Given a research description, SurveyMaster (i) calibrates generation with a discipline-specific writing skill, (ii) builds a comprehensive paper pool through seed-paper-driven hierarchical retrieval across a 160-million-paper scientific database, and (iii) anchors the manuscript on a small set of core papers that form its conceptual backbone. To evaluate SurveyMaster, we introduce SciSurveyBench, a multidisciplinary benchmark of 100 expert-curated synthesis tasks across ten natural- and social-science disciplines. SurveyMaster achieves the highest overall score in all ten disciplines. It matches or exceeds expert-authored reviews on topical relevance, coverage, coherence and critical analysis, and improves citation grounding from 3.63 to 4.84 (fact correctness) and from 3.62 to 4.49 (citation precision) over the strongest automatic baseline. Controlled ablations confirm that these gains come from the three design choices, not from prompt-level tuning. By recasting survey generation as evidence-grounded cognitive distillation, SurveyMaster provides a scalable framework for evidence-grounded literature synthesis.

[1]Shanghai Jiao Tong University, Shanghai, China [2]SciLand, Shanghai, China. Correspondence to: Siheng Chen <sihengc@sjtu.edu.cn>.

Accepted at the AI for Science Workshop at ICML 2026.

1. Introduction

Scientific progress requires the rigorous synthesis of existing knowledge. Literature reviews are the primary means of achieving this synthesis (Snyder, 2019; Grant & Booth, 2009; Bolanos et al., 2024). Ideally, reviews perform cognitive distillation: they transform heterogeneous evidence into a structured narrative, identify foundational work, highlight critical uncertainties, and ground conclusions in precise citations. Yet the scale of the literature is outpacing the expert labor required to synthesize it. Global science and engineering output reached 3.3 million articles in 2023 (National Center for Science and Engineering Statistics, 2025), whereas high-quality evidence synthesis often requires months to more than a year of coordinated expert work. This growing mismatch between publication growth and human synthesis capacity has created a structural bottleneck, slowing timely knowledge integration and impeding cross-disciplinary exchange across increasingly fragmented fields (Wang et al., 2023b; Chu & Evans, 2021; Ofori-Boateng et al., 2024).

Recent automatic literature review systems powered by large language models (LLMs) have dramatically improved efficiency, producing survey-like text in hours (Wang et al., 2024; Yan et al., 2025; Liang et al., 2025), but they still fall short of cognitive distillation. Three requirements make cognitive distillation difficult to scale. First, it demands domain-specific calibration to each field's evidentiary standards and conceptual vocabulary, so that domain concepts are used in their precise disciplinary sense rather than as generic terms inside fluent prose (Zhang et al., 2024; Wang et al., 2023a). Second, it requires a comprehensive literature pool that traces the full intellectual lineage of a topic without overweighting popular subareas or omitting foundational work (Kandpal et al., 2023; Agarwal et al., 2024). The exponential growth and fragmentation of the global literature makes this kind of comprehensive coverage progressively harder to achieve. Third, it relies on argumentative logic to trace how studies corroborate, conflict with or build on one another, turning heterogeneous evidence into a coherent cumulative narrative (Hoang & Kan, 2010; Hu & Wan, 2014; Zhu et al., 2023; Hu et al., 2024). Together these requirements impose a workload that already exceeds what expert

reviewers alone can keep up with at the scale of today's literature.

Here we present SurveyMaster, an artificial intelligence system designed to generate expert-level literature reviews across multiple scientific disciplines. Given a target research description as input, SurveyMaster breaks the complex task of cognitive distillation into three clear analytical steps that mirror expert practice. It begins with domain-specific calibration, extracting a field-specific writing skill from a pre-built skill library and dynamically aligning its generation with precise disciplinary conventions. It then builds a comprehensive paper pool by retrieving expert-curated seed surveys relevant to the target topic, using them to generate search queries, and performing hierarchical retrieval across a global scientific database. Finally, it enforces argumentative logic by first identifying a core subset of papers that form the review's conceptual backbone and then organizing the remaining literature around this backbone to construct a cumulative scientific narrative.

To evaluate whether SurveyMaster delivers on cognitive distillation, we introduce SciSurveyBench, a multidisciplinary benchmark of 100 synthesis tasks across ten natural- and social-science disciplines, evaluated with a multidimensional protocol covering outline quality, content quality and citation grounding. The citation-grounding axis directly penalizes the citation hallucinations that have become the dominant failure mode of LLM-generated scientific text (Tang et al., 2025).

On SciSurveyBench, SurveyMaster achieves the highest overall score in every one of the ten disciplines. It matches or exceeds expert-authored reviews on topical relevance, coverage, coherence and critical analysis, scoring 4.06/5 against 3.45/5 on critical analysis, and trails by only 0.13 points on instruction adherence and 0.31 points on content relevance. SurveyMaster also raises fact correctness from 3.63 to 4.84 and citation precision from 3.62 to 4.49 over the strongest automatic baseline. Controlled ablations further show that these gains come from the three structural choices outlined above, not from prompt-level tuning. When the discipline-specific writing skill is replaced with one from another field, the survey starts using concepts and vocabulary from that other field. When seed surveys are removed, the retrieval drifts toward loosely related topics. When core papers are removed, the manuscript's instruction adherence and content relevance fall. In summary, these results support cognitive distillation as a practical design for expert-level literature synthesis. They also suggest a practical way for the scientific community to keep pace with the expanding literature.

2. Methods

2.1. Problem Formulation

SurveyMaster is an LLM-powered system (θ) that turns a user-specified research description q into a structured literature review grounded in an explicitly constructed evidence base. Given q, the system produces five outputs: (i) the writing skill σ calibrates generation to the target discipline; (ii) the curated paper pool $\mathcal{P}$ holds the available evidence; (iii) the outline O is anchored on a small set of core papers that form the conceptual backbone of the review; (iv) the section-level evidence sets $\{\mathcal{C}_k\}_{k=1}^{K}$ map papers to outline sections; and (v) the survey manuscript $\mathcal{S} = \{s_1, \ldots, s_K\}$ is the final generated review.

$$(\sigma,\, \mathcal{P},\, O,\, \{\mathcal{C}_k\}_{k=1}^{K},\, \mathcal{S}) \;=\; \texttt{SurveyMaster}_{\theta}(q). \quad (1)$$

This formulation (1) treats review generation as evidence-grounded synthesis rather than one-pass text production. The skill shapes writing style, the paper pool defines the available evidence, the core papers form the conceptual backbone, and the section-level evidence sets constrain local writing decisions.

2.2. SurveyMaster: Literature Review System

SurveyMaster realizes the three analytical steps of cognitive distillation introduced in Section 1 through four stages that follow the workflow in Figure 1.

Preparation. The Preparation stage carries out the domain-specific calibration step. SurveyMaster reads the input query, identifies the target discipline and topic, and retrieves the corresponding writing skill from a pre-built skill library that covers 14 disciplines. The skill provides discipline-specific guidance on exposition style, section organization and figure roles. The library was constructed offline. For each discipline, we collected representative human-written reviews and summarized their shared writing conventions with an LLM, and a domain expert then verified or revised the result.

Seed-driven hierarchical retrieval. This stage carries out the comprehensive evidence assembly step. SurveyMaster first retrieves a small set of authoritative seed surveys from the multidisciplinary literature database via the PasaMaster search engine (Du et al., 2026). An LLM then converts these seeds into a hierarchical research outline. The survey topic is partitioned into a small number of thematic research aspects, and each aspect is expanded into specific keyword expressions (Figure 1). Papers are retrieved against this expanded keyword set. Each retrieved paper passes a relevance filter based on the cosine similarity between its abstract embedding and the survey topic, and is then deduplicated and metadata-completed before entering the pool. Because the keyword set is structured around thematic aspects rather than a single flat keyword list, retrieval covers

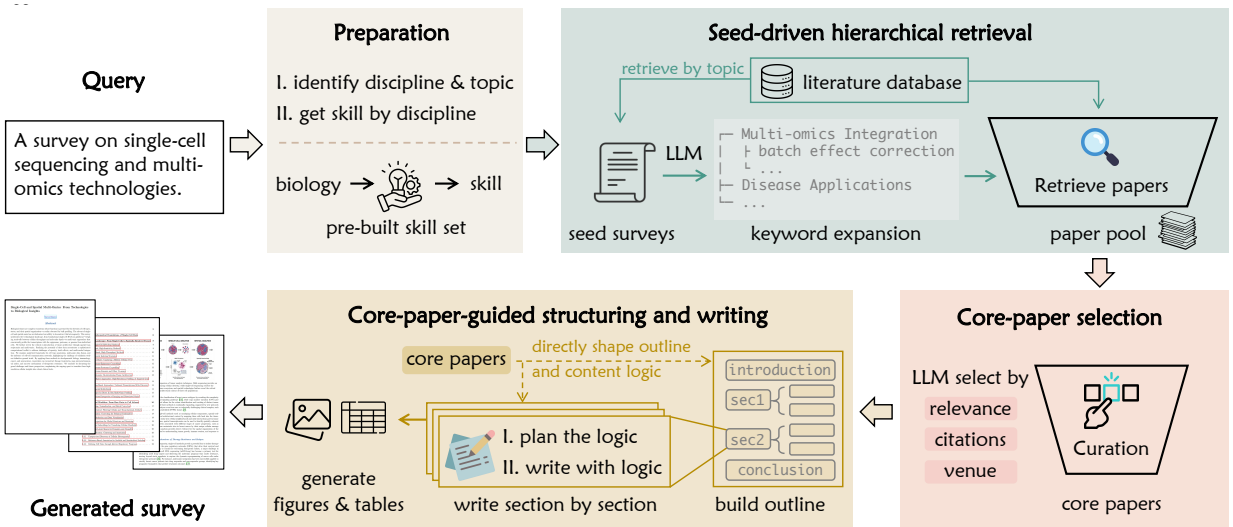

Figure 1. Overview of SurveyMaster for expert-level literature review generation.

topic-defining sub-areas of each discipline rather than drifting toward broader or loosely related topics.

Core-paper selection. This stage extracts the conceptual backbone of the review. From the paper pool, an LLM selects a small set of core papers (typically less than 5% of the pool) and organizes them into a few thesis-bearing buckets. Each bucket carries an explicit thesis statement that summarizes its intellectual claim, and each core paper inside a bucket is tagged with its role, such as milestone, method, foundation, theory or survey. The selection step is conditioned on the paper-level metadata returned by the search engine, such as venue, citation count and impact factor. This means that the LLM evaluates citation impact and venue quality from explicit metadata, not from its parameter knowledge. Because the candidate pool can exceed the LLM's context window, selection is performed in two stages. The LLM first proposes the initial buckets and core papers from the highest-cited candidates. Subsequent batches of candidates are then merged into the running bucket structure, with additions, replacements and re-rankings applied as new evidence arrives.

Core-paper-guided structuring and writing. This stage anchors the manuscript to the bucket structure produced above. SurveyMaster reorganizes an initial outline so that its sections correspond to the thesis-bearing buckets. A discipline style-compliance check is then applied; if the outline does not satisfy the discipline's writing skill, a single regeneration is triggered. Each remaining paper in the pool is then assigned by an LLM to one or more sections of the outline. This step follows the section-evidence assignment scheme of SurveyX (Liang et al., 2025) and produces the section-level evidence sets $\{\mathcal{C}_k\}_{k=1}^{K}$ introduced above. The manuscript is then written section by section. Each section is generated conditioned on its assigned evidence set, the thesis of its corresponding bucket, and all previously written sections. This conditioning lets the manuscript read as one cumulative argument rather than a stack of locally retrieved summaries. After the main body and abstract are produced, a sanitization pass cleans the output. It removes any citation whose key does not match a paper in the pool. This last step provides a final safeguard against citation hallucinations.

2.3. SciSurveyBench

Benchmark construction. To evaluate SurveyMaster on multidisciplinary literature synthesis, we construct SciSurveyBench, comprising 100 independent survey generation tasks. We select 10 scientific disciplines spanning both natural and social sciences, including biology, chemistry, physics, earth science, computer science, sociology, economics, psychology, medicine and civil engineering, to ensure broad coverage. Within each discipline, human experts manually curated 10 highly cited, peer-reviewed review articles published since 2021. Topical overlap was minimized to preserve task diversity and to avoid over-representation of closely related subfields. For each reference review, we then reverse-engineered a simulated user query that captures the central scope and intent of the target survey. This procedure yielded 100 standardized queries paired with 100 expert-written surveys.

Evaluation protocol. Evaluating full-length, expert-level scientific surveys requires moving beyond generic text-quality metrics. We developed an automated evaluation protocol leveraging an LLM-as-a-Judge (Gu et al., 2024), which assesses the generated manuscripts across eight fine-grained dimensions grouped into three primary axes:

1) **Outline quality** measures the structural foundation of

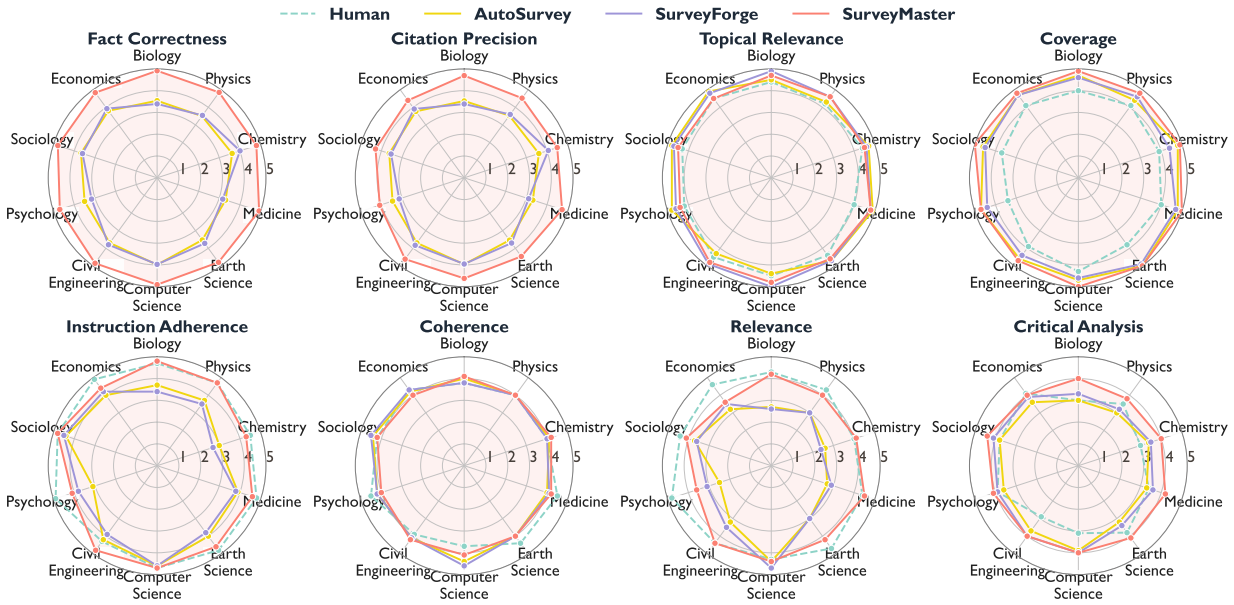

Figure 2. Multi-dimensional benchmark comparison across methods. Each radar plot reports one evaluation metric across the ten disciplines, with each radial axis corresponding to one discipline and each value averaged over ten benchmark tasks. Scores range from 1 to 5, with higher values indicating better performance.

the survey, including *Topical Relevance*, which evaluates alignment with the input query, and *Coverage*, which measures the breadth of the intellectual lineage represented. 2) **Content quality** measures the analytical and discursive quality of the manuscript, including *Instruction Adherence*, *Coherence*, *Relevance* and *Critical Analysis*. *Critical Analysis* in particular evaluates whether the survey constructs argumentative links across studies rather than merely summarizing them in isolation. 3) **Citation quality** assesses factual grounding and explicitly penalizes citation hallucinations, a major failure mode in scientific text generation. We quantify this axis using two metrics. *Fact Correctness* is the proportion of cited claims supported by at least one associated reference,

$$\text{FactCorrectness} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\exists r \in R_i,\ \text{Support}(r, c_i) = 1]\,,$$

where N is the number of extracted cited claims, c_i is the i-th claim, and R_i is its associated reference set. *Citation Precision* is the proportion of faithful claim-reference pairs among all evaluated pairs,

$$\text{CitationPrecision} = \frac{\sum_{i=1}^{N}\sum_{r\in R_i} \mathbf{1}[\text{Support}(r, c_i) = 1]}{\sum_{i=1}^{N} |R_i|},$$

where the denominator is total number of claim-reference pairs and the numerator counts those judged faithful.

We use GPT-5.4 (OpenAI, 2026) as the judge model for outline and content quality, and Gemini-3-Flash (Google DeepMind) for citation-quality evaluation. The judge temperature is set to 0.

3. Experimental Results

3.1. Experimental Setup

We compared SurveyMaster against two representative automatic survey-generation methods, AutoSurvey (Wang et al., 2024) and SurveyForge (Yan et al., 2025). To ensure a fair comparison, all three systems used Gemini-2.5-Pro (Comanici et al., 2025) as the backbone LLM for survey generation. For AutoSurvey and SurveyForge, we used their official implementations and default literature databases. We use the 100 collected human-written surveys as expert reference baselines, and the corresponding 100 simulated queries as standardized inputs for all automated systems. The construction of SciSurveyBench and the LLM-as-a-Judge evaluation protocol used throughout this section are described in Methods (Section 2.3). Additional implementation details are provided in Appendix A.1.

3.2. Main Results

SurveyMaster substantially outperforms automatic baselines on the content metrics that depend most on the retrieved literature. Figure 2 summarizes performance across the eight evaluation criteria and ten disciplines. At the outline level, SurveyMaster achieves the strongest automatic score on coverage, improving over the stronger baseline by 0.18 points on average, while remaining competitive on topi-

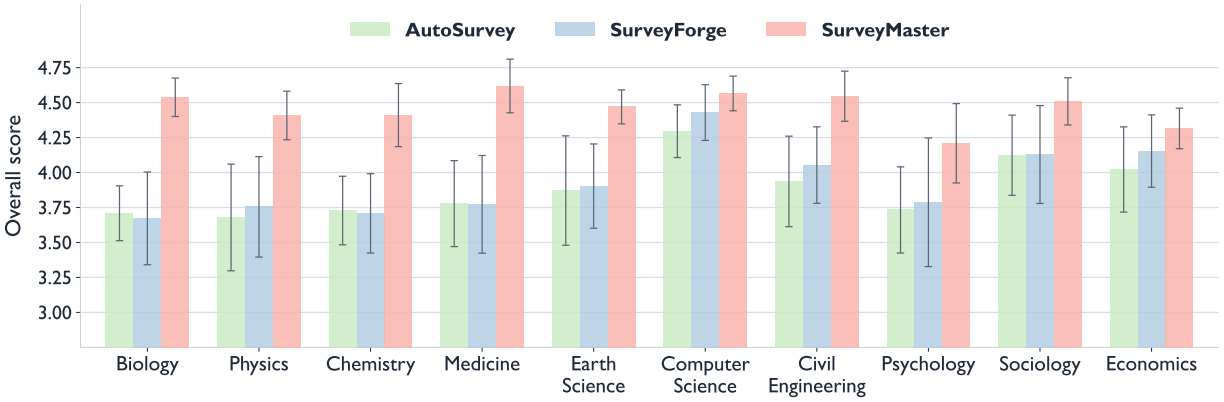

Figure 3. Overall benchmark performance across disciplines. Bars show the mean overall score for each automatic method in each discipline, computed as the unweighted average of the eight evaluation metrics. Error bars denote the standard deviation across the ten benchmark tasks within each discipline.

cal relevance. The larger separation appears in content-level criteria: SurveyMaster improves instruction adherence by 0.72 points, relevance by 0.88 points and critical analysis by 0.37 points over the strongest automatic baseline, while coherence is comparable across the strongest systems. This shows that the advantage does not come from more fluent local writing but from better preserving the requested scope and organizing the evidence into a synthesis that is tailored to the requested topic. Two design choices most plausibly account for this pattern: seed-driven hierarchical retrieval, which keeps the paper pool aligned with the specific subareas that define each topic, and the core-paper backbone, which gives each section a small set of conceptual anchors so that the writing does not slide onto loosely related topics.

SurveyMaster matches or exceeds the human-authored reviews on most outline and content metrics. SurveyMaster matches or exceeds the human-authored reviews on topical relevance, coverage, coherence and critical analysis on average, and trails by 0.13 points on instruction adherence and 0.31 points on content relevance (Figure 2). The largest gap is on coverage, where the human-authored reviews score 0.96 points below SurveyMaster and lower than every automatic system. We do not read this as a deficiency of the human authors: published reviews are written for a specific editorial purpose, and authors deliberately narrow down the literature they include, whereas an automated system queried with the same topic is asked to produce a comprehensive synthesis and scores higher for the kind of breadth that targeted human reviews do not aim for. We do not report fact correctness or citation precision for the human-authored reviews, because verifying every claim against the full text of every cited paper is rarely feasible at the scale of a multidisciplinary benchmark, where many papers sit behind paywalls or appear as scanned PDFs without searchable text.

SurveyMaster consistently improves citation grounding over the strongest automatic baseline. Citation grounding measures whether the claims a manuscript makes are actually supported by the papers it cites, a requirement that distinguishes scientific synthesis from plausible long-form text. SurveyMaster improves fact correctness by 1.21 points and citation precision by 0.87 points over the strongest automatic baseline (Figure 2). These margins are substantial because citation errors are not always visible from fluency or topical relevance alone: a manuscript can read coherently while attaching its claims to weakly related or unsupported papers. Three design choices in SurveyMaster directly target this kind of citation error: every citable reference is restricted to the curated paper pool; each section is written from a section-specific evidence set, not from the model's prior knowledge; and the central claims of each section are organized around a small number of core papers whose roles make them the natural papers to cite. A final sanitization pass then removes any citation whose key is not in the paper pool, catching the residual hallucinated references that slip past the previous steps.

SurveyMaster achieves the highest overall score in every one of the ten disciplines. Aggregating the eight metrics into an overall score for each discipline (Figure 3) confirms that this advantage is not confined to a single favorable field or metric group. The improvement is most pronounced in several natural-science, medical and engineering disciplines, whereas the gains are smaller but still consistent in fields where the baselines are already comparatively strong. This broad pattern supports the motivation introduced above: scalable review generation requires more than producing survey-like prose. It requires cognitive distillation, in which a system constructs a comprehensive evidence base, preserves disciplinary scope and turns the literature into a grounded cumulative narrative.

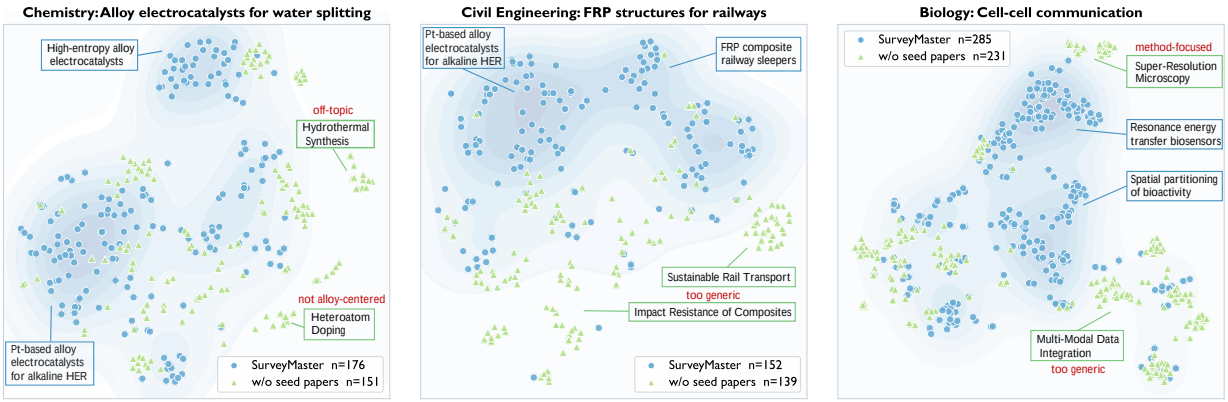

Figure 4. Ablation of seed-paper-driven retrieval. Each point represents a retrieved paper, embedded by its abstract and projected with t-SNE. Blue circles indicate SurveyMaster results, and green triangles indicate retrieval without seed-paper guidance.

Table 1. Ablation of core-paper-guided organization. Performance comparison between SurveyMaster and SurveyMaster without core papers on the 100 SciSurveyBench tasks. Each cell shows mean ± standard deviation across tasks. Reference metrics are scaled to a 5-point score for comparability with the content metrics.

Method	Topical Relevance	Coverage	Coherence	Relevance	Critical Analysis	Instruction Adherence	Fact Correctness	Citation Precision	Average
SurveyMaster	4.62±0.49	4.88±0.33	4.10±0.30	4.11±0.58	4.06±0.37	4.58±0.54	4.84±0.12	4.49±0.31	4.46
SurveyMaster w/o core papers	4.54±0.50	4.82±0.39	4.05±0.30	3.85±0.73	4.00±0.40	4.25±0.83	4.84±0.12	4.58±0.22	4.37
Change	↓ 0.08	↓ 0.06	↓ 0.05	↓ 0.26	↓ 0.06	↓ 0.33	0.00	↑ 0.09	↓ 0.09

3.3. Ablation of Seed-paper-driven Retrieval

Here we investigate the effect of seed-paper-driven query expansion on the construction of the paper pool. In the ablated system, the LLM generates retrieval keywords directly from the user query, without first reading seed surveys; all downstream retrieval and filtering steps are unchanged. We compare SurveyMaster with this no-seed baseline on three representative topics: alloy electrocatalysts for water splitting, fiber-reinforced polymer (FRP) structures for railways, and cell-cell communication. For each topic, we project retrieved-paper abstracts with t-SNE and ask whether the resulting pool covers the core subtopics of the query (Figure 4).

The comparison in Figure 4 shows two consistent effects of seed-paper guidance. First, the blue regions mark topic-relevant clusters retrieved by SurveyMaster but largely missed by the no-seed baseline, including high-entropy and Pt-based alloy electrocatalysts for alkaline HER (chemistry), FRP railway sleepers (civil engineering) and resonance-energy-transfer biosensors together with spatial bioactivity partitioning (biology). Second, the green regions mark clusters enriched in the no-seed baseline that are related but less central to the target scope: hydrothermal synthesis as a broad preparation route rather than an alloy-centered electrocatalyst topic, sustainable rail transport as a system-level theme rather than an FRP-structure theme, and multi-modal data integration as a generic methodological topic rather than a mechanistic cell-cell communication theme. Seed surveys therefore improve comprehensiveness by preserving coverage of topic-specific subareas while reducing drift toward broad or superficial concepts.

3.4. Ablation of Core-paper-guided Organization

The two metrics that exhibit notable declines are instruction adherence at -0.33 and content relevance at -0.26. Both metrics assess the degree to which the manuscript remains aligned with the requested scope. This is consistent with the role we ascribed to core papers, which act as conceptual anchors that constrain what each section is supposed to argue. Without them, the writer reasons over a large pool of related-but-not-central papers and is more likely to drift toward generic methodological or background material. Outline metrics, coherence and critical analysis change less, as expected. The paper pool itself is unchanged, so a reasonable outline and locally fluent prose are still recoverable. The loss is concentrated in section-level argumentation.

The two reference-grounding metrics behave differently from the scope metrics. Fact correctness is essentially unchanged (4.84 versus 4.84), and citation precision is even slightly higher for the no-core variant (4.58 versus 4.49). The small size of both gaps suggests that the core-paper backbone does not influence citation faithfulness. Instead,

this accuracy stems from two design choices. First, all citable references are strictly limited to the curated paper pool. Second, each section is generated using a specific set of evidence rather than relying on the model's prior knowledge. The small rise in pair-level precision for the no-core variant partly reflects that it attaches fewer references per claim and reuses a smaller set of papers, which removes weakly aligned claim-reference pairs. Core papers therefore act mainly as scope anchors, and their contribution is concentrated in instruction adherence and content relevance rather than in raw citation grounding.

4. Conclusion

We present SurveyMaster, an agentic AI system that performs cognitive distillation by calibrating writing to each discipline, assembling a paper pool from seed surveys and anchoring the manuscript on a small set of core papers. Across all ten natural- and social-science disciplines, SurveyMaster matches or exceeds human-authored reviews on most outline and content metrics and substantially improves citation grounding over the strongest automatic baseline, indicating a general design principle rather than field-specific tuning. With cognitive distillation now achievable at scale, keeping pace with an exponentially growing literature ceases to be a structural barrier to science and becomes a question of how the scientific community chooses to deploy these systems.

Acknowledgements

The authors thank the developers of the 160-million-paper literature database at DP Technology. This work was supported by the Shanghai Jiao Tong University–DP Technology Super Scientific Research Platform Strategic Cooperation Agreement (AI-X Research Institute).

References

Agarwal, S., Sahu, G., Puri, A., Laradji, I. H., Dvijotham, K. D., Stanley, J., Charlin, L., and Pal, C. Llms for literature review: Are we there yet? *arXiv e-prints*, pp. arXiv–2412, 2024.

Bolanos, F., Salatino, A., Osborne, F., and Motta, E. Artificial intelligence for literature reviews: opportunities and challenges: F. bolaños et al. *Artificial Intelligence Review*, 57(10):259, 2024.

Chu, J. S. G. and Evans, J. A. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41), October 2021. ISSN 1091-6490. doi: 10.1073/pnas.2021636118.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Du, Y., Jin, T., Kang, J., Pang, X., Chai, J., Miao, T., Liu, F., Wang, W., Yao, S., Zhang, Y., and Chen, S. Towards self-evolving agentic literature retrieval, 2026. URL https://arxiv.org/abs/2605.14306.

Google DeepMind. Gemini 3 flash. https://deepmind.google/models/gemini/flash/. Accessed: 2026-04-08.

Grant, M. J. and Booth, A. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, 26(2):91–108, 2009. ISSN 1471-1842. doi: 10.1111/j.1471-1842.2009.00848.x.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Hoang, C. D. V. and Kan, M.-Y. Towards automated related work summarization. In *Coling 2010: Posters*, pp. 427–435, Beijing, China, 2010. Coling 2010 Organizing Committee. URL https://aclanthology.org/C10-2049/.

Hu, Y. and Wan, X. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1624–1633, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1170. URL https://aclanthology.org/D14-1170/.

Hu, Y., Li, Z., Zhang, Z., Ling, C., Kanjiani, R., Zhao, B., and Zhao, L. Hireview: Hierarchical taxonomy-driven automatic literature review generation, 2024. URL https://arxiv.org/abs/2410.03761.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. Large language models struggle to learn long-tail knowledge, 2023. URL https://arxiv.org/abs/2211.08411.

Liang, X., Yang, J., Wang, Y., Tang, C., Zheng, Z., Song, S., Lin, Z., Yang, Y., Niu, S., Wang, H., et al. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

National Center for Science and Engineering Statistics. Publication output by geography and scientific field. Science and Engineering Indicators 2026, National Science Foundation, 2025. URL https://ncses.nsf.gov/pubs/nsb20257/publication-output-by-geo

graphy-and-scientific-field. Accessed 26 April 2026.

Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., and Moreno-Garcia, C. F. Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial intelligence review*, 57(8):200, 2024.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026. Accessed: 2026-04-08.

Snyder, H. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339, 2019. ISSN 0148-2963. doi: 10.1016/j.jbusres.2019.07.039.

Tang, X., Duan, X., and Cai, Z. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1602–1617, 2025.

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., and Zhang, Y. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity, 2023a. URL https://arxiv.org/abs/2310.07521.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C. W., Bengio, Y., and Zitnik, M. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023b. ISSN 1476-4687. doi: 10.1038/s41586-023-06221-2.

Wang, Y., Guo, Q., Yao, W., Zhang, H., Zhang, X., Wu, Z., Zhang, M., Dai, X., Zhang, M., Wen, Q., Ye, W., Zhang, S., and Zhang, Y. Autosurvey: Large language models can automatically write surveys, 2024. URL http://arxiv.org/abs/2406.10252.

Yan, X., Feng, S., Yuan, J., Xia, R., Wang, B., Zhang, B., and Bai, L. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing, 2025. URL https://arxiv.org/abs/2503.04629.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., and Han, J. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8783–8817, 2024.

Zhu, K., Feng, X., Feng, X., Wu, Y., and Qin, B. Hierarchical catalogue generation for literature review: A benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6790–6804, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.453. URL https://aclanthology.org/2023.findings-emnlp.453/.

A. Experimental Results Details

A.1. Baselines Setup

For AutoSurvey and SurveyForge, we used the authors' official implementations and retained their provided literature databases. The baseline sources were: AutoSurvey, `github.com/AutoSurveys/AutoSurvey`; and SurveyForge, `github.com/InternScience/SurveyForge`. To adapt the baselines to SciSurveyBench, we made only minimal interface-level changes needed to pass our standardized user queries and collect outputs in a common format. We did not alter the core retrieval, planning or generation logic of either baseline.

A.2. Domain-skill ablation: qualitative case study

This appendix presents a qualitative case study of the domain-skill ablation, complementing the quantitative ablations reported in the main paper. Using a sociology topic, we hold the query and retrieved evidence fixed and condition generation on three writing skills: a correct-discipline skill (*Sociology*), an empty no-skill condition, and a wrong-discipline skill (*Chemistry*). The excerpts in Figure 5 are taken directly from the generated surveys.

This example shows that the domain skill mainly affects the conceptual framing of the generated survey. With the Sociology skill, the survey describes digital inequity using social-science concepts such as stratification, intersectionality and compounded marginalization. Without a domain skill, the output remains broadly relevant but uses more generic language about opportunity and inclusion. Under the Chemistry skill, the generation imports reaction-mechanism and catalysis metaphors, introducing inappropriate chemistry metaphors into the sociological analysis. This explains why the wrong-skill condition can preserve surface topic coverage while degrading relevance and instruction adherence.

Query

Using the systematic integrative review evidence, summarize what is known about digital inequity in developed countries and which populations are most often studied as vulnerable. Explain the review's aims, its four synthesized dimensions—digital literacy, affordability, equity-deserving-group-sensitive content, and infrastructure availability/quality—and the major barriers involving internet access, cost, skills, and language. Close with the review's intersectionality framing and policy implications.

Correct skill: Sociology

Verbatim excerpt.

"In an era of accelerating digitalization, digital inequity has emerged as a **primary axis of social stratification**, threatening to amplify existing social cleavages in developed countries."

"Adhering to PRISMA-style reporting and employing an **intersectionality framework**, the review analyzes how overlapping social identities produce **compounded forms of marginalization**."

Why this is good. The sociology skill keeps the survey anchored to social stratification, vulnerable populations, intersectionality, digital exclusion, and policy-facing inequity.

Empty skill: Generic

Verbatim excerpt.

"In an era where digital participation is fundamental to **economic opportunity, civic engagement, and social inclusion**, the gap between the digitally connected and the excluded has become a primary axis of modern inequality."

"This systematic integrative review, adhering to PRISMA reporting standards, addresses the **enduring challenge of digital inequity** within developed countries."

Why this is weaker. The generic skill remains relevant, but gives weaker disciplinary guidance. It is less explicitly organized around compounded disadvantage and intersectional marginalization.

Wrong skill: Chemistry

Verbatim excerpt.

"The digital revolution is frequently presented as a **catalyst** for unprecedented societal progress; however, like a powerful but poorly selective **catalytic system**, its deployment has yielded significant social inequity as a primary, **toxic byproduct**."

"This systematic integrative review aims to elucidate the **reaction mechanism** of digital inequity within developed countries."

Why this is bad. The chemistry skill contaminates a sociology review with reaction-mechanism and catalysis metaphors, weakening the requested social-science framing.

Observed score pattern. For this non-CS case, the correct-discipline skill obtains the strongest content score (**C Avg.**=4.25), followed by the empty-skill condition (4.00), while the chemistry wrong-discipline skill drops sharply (2.75), mainly because mismatched disciplinary framing degrades relevance and instruction adherence.

Figure 5. **Qualitative case study of domain-skill ablation on a sociology topic.** The excerpts are copied from generated surveys for the same digital-inequity query under correct, empty, and wrong writing-skill conditions. The correct Sociology skill preserves the intended social-science framing; the empty skill remains broadly relevant but less disciplinary; the wrong Chemistry skill introduces reaction-mechanism and catalysis language that contaminates the sociological analysis.