

Artificial Intelligence for Scientific Discovery: A Focused Review of Large Language Models for Scientific Literature, Table Understanding, and Numerical Reasoning

Swara Sandip Lende, Prof. Bangar Abhishek Popat, Prof. Bhosale S.B.

Dept. of Computer Engineering
Jaihind College of engineering, Pune, India

Abstract — Artificial intelligence (AI) is increasingly used to accelerate scientific discovery, from predicting protein structures and molecular properties to extracting knowledge from the scientific literature. A large share of experimental evidence in papers is reported in numerical tables, yet automated analysis of such tables remains difficult. This paper presents a focused narrative review of verified sources on AI for scientific discovery, concentrating on large language model (LLM) based understanding of scientific text, scientific tables and numerical reasoning. It examines three connected themes: knowledge extraction and question answering over scientific text, table question answering benchmarks, and reasoning methods (direct prompting, chain-of-thought, program-of-thought, self-consistency and graph-based reasoning). The literature shows that scientific table benchmarks have moved from mostly lookup tasks towards a 1,607-question benchmark that requires multi-step numerical reasoning; on it, GPT-4.1 mini reached only 27.0 exact match under direct prompting on the computation-and-language subset, chain-of-thought prompting raised this to 71.5, and accuracy still fell from 74.6% on small tables to 59.2% on large ones. However, the evidence is concentrated in computer science tables, depends on LLM-generated and LLM-filtered questions, and agreement among human verifiers was only moderate (Fleiss' kappa 0.55). The review contributes a comparative analysis of benchmarks and methods, a proposed conceptual architecture linking extraction, cell selection, stepwise reasoning, verification and expert review, and a prioritised list of research gaps.

Keywords— AI for Scientific Discovery, Large Language Models, Scientific Table Question Answering, Numerical Reasoning, Benchmarks, Chain-of-Thought.

I. INTRODUCTION

AI is playing a growing role in accelerating scientific knowledge discovery. Recent work discusses the progress, opportunities and challenges of generative AI for scientific discovery [1], applies LLMs to molecular property prediction [2], and, in a widely known example, predicts protein structures with high accuracy [3]. Much of the supporting analysis work, however, starts from the scientific literature, where experimental results and statistical analyses are reported in tables [4].

Scientific tables often contain information that is not stated in the surrounding text, and using them requires discerning patterns, trends and quantitative relationships [4]. Real tasks such as aggregating results across experimental conditions, analysing trends and comparing metrics need multi-step numerical reasoning, not only retrieval of single values [4]. This review therefore examines three strands together: (i) knowledge extraction and question answering over scientific

text, (ii) scientific table question answering benchmarks, and (iii) reasoning methods applied to them. It asks how mature the evidence is and proposes a conceptual architecture that addresses the weaknesses found.

II. REVIEW METHODOLOGY

This is a focused narrative review, not a systematic one. The main source is a recent benchmark paper on numerical reasoning over scientific tables, SciNumQA [4], together with the studies it cites. Descriptions of related benchmarks and methods, and all reported numbers, are taken as stated in [4] and in its tables and were not independently re-extracted from the original papers. For the works in [1], [2] and [5]–[8], only their scope as given by their titles and by [4] is used. Two standard sources, [3] and [15], were added for context. No separate database search was run and no study counts are claimed. The review is therefore narrow and is not a full survey of AI for scientific discovery; this is stated as a limitation in Section VIII.

III. KNOWLEDGE EXTRACTION AND QUESTION ANSWERING OVER SCIENTIFIC TEXT

Earlier AI work for scientific knowledge has focused on entity and relation extraction and on question answering over the scientific literature [4]. SciER is an entity and relation extraction dataset for datasets, methods and tasks in scientific documents [5].

Luan et al. addressed multi-task identification of entities, relations and coreference for scientific knowledge graph construction [6]. SciBench evaluates college-level scientific problem-solving abilities of LLMs [7], and Xu et al. studied multi-hop science question answering using reasoning chains [8]. These resources target text. Tables, which carry most of the numerical evidence, are the subject of the next section.

IV. SCIENTIFIC TABLE QUESTION ANSWERING

1. General-domain and finance benchmarks

WikiTableQuestions [11] covers the general domain and tends to emphasise entity lookup or simple relational queries, while DocMath-Eval [12] covers finance and mainly involves transaction-based or arithmetic reasoning [4]. Neither reflects the multi-metric, comparative analyses typical of scientific tables [4].

2. Scientific table benchmarks

SciTabQA [9] and SCITAT [10] derive questions from scientific papers. According to [4], SciTabQA is limited in scale and most of its questions are simple cell lookup or aggregation. SCITAT covers both tables and text, but only about half of its questions involve explicit numerical reasoning, and both datasets contain many small tables that lead to trivial questions [4]. SciGen, a dataset for reasoning-aware text generation from scientific tables [13], supplies tables from computer science papers that were reused as the source for SciNumQA [4].

3. SciNumQA construction

SciNumQA contains 1,607 question-answer pairs over 572 scientific tables [4]. Tables of 3×3 or smaller, and those with two or fewer effective numerical rows or columns, were removed. Eight reasoning types were defined, following earlier work [9], [14]: cell comparison, condition comparison, difference calculation, ratio calculation, change analysis, trend comparison, average or sum calculation, and minimum or maximum identification.

An LLM generated 5,944 candidate pairs, LLM-based filtering for correctness, relevance and insightfulness kept 2,064, and four undergraduate computer science students verified the final 1,607 pairs. Agreement on a 100-pair subset was moderate, with Fleiss' kappa of 0.55 [4], [15].

V. REASONING METHODS AND REPORTED PERFORMANCE

The benchmark was used to evaluate Qwen-2.5-Instruct (7B, 72B), LLaMA-3.2-Instruct (3B, 11B, 90B), GPT-4.1 (mini, nano) and GPT-o4-mini [4]. All models were tested with direct prompting (DP). Further reasoning methods, applied only to GPT-4.1 mini and GPT-4.1 nano, were chain-of-thought (CoT) [16], program-of-thought (PoT) [17], MIX-SC [18] and GraphOTTER [19]. Metrics were exact match (EM) and an LLM-as-judge score [4].

Under DP, models struggled: GPT-4.1 mini, the strongest, reached 27.0 EM and 32.8 LLM score on the computation-and-language (C&L) subset [4].

Explicit reasoning helped: CoT raised GPT-4.1 mini to 71.5 EM and 82.7 LLM score on C&L, whereas PoT, MIX-SC and GraphOTTER reached 43.3, 69.3 and 49.1 EM, so more complex methods did not beat CoT [4]. Performance also fell with table complexity.

On the Other-domain test set, GPT-4.1 mini with CoT dropped from 74.6% EM on small tables to 59.2% on large ones, and from 72.5% to 60.6% on tables with hierarchical headers; GraphOTTER, which selects only relevant cells, showed a slight increase on large tables [4].

The authors' error analysis found failures even when all needed values were present: reading a wrong value (22% instead of 10% for one metric) and selecting the wrong metric (Label Error instead of ECE) in a multi-step aggregation [4]. These are errors of cell and metric selection, not only of arithmetic.

VI. COMPARATIVE ANALYSIS

Table I compares the benchmarks and Table II the reasoning methods, using figures reported in [4]. Scores are not comparable across benchmarks because tasks and metrics differ.

Table 1: Comparison of Table Question Answering Resources

Resource	Domain / content	Scale	Key characteristic	Limitation (as stated in [4])
WikiTableQuestions [11]	General-domain tables	Not given in [4]	Entity lookup, simple relational queries	Not scientific; little complex numerical reasoning
DocMath-Eval [12]	Finance documents	Not given in [4]	Transaction-based or arithmetic reasoning	Does not reflect multi-metric scientific analysis
SciTabQA [9]	Scientific tables	34 complex numerical tables; 199 QA pairs	Mostly cell lookup and basic comparison	Limited scale; many small tables
SCITAT [10]	Scientific tables and text	445 tables; 438 QA pairs	More varied reasoning types, with textual context	About half need explicit numerical reasoning
SciGen [13]	Tables from computer science papers	Source tables for [4]	C&L and Others domains	Computer science only
SciNumQA [4]	Scientific tables, short-form QA	572 tables; 1,607 QA pairs	Eight numerical reasoning types; multi-step	CS-centred; LLM generation bias; kappa 0.55

Descriptions and counts are as reported in [4] and were not re-extracted from the original papers.

Table 2: Reasoning Methods on Scinumqa (Gpt-4.1 Mini; Exact Match and Llm Score)

GPT-4.1 mini, method	C&L EM	C&L LLM score	Other EM	Other LLM score
Direct prompting	27.0	32.8	33.1	39.2
+ CoT	71.5	82.7	66.4	80.0
+ PoT	43.3	50.9	40.8	51.8
+ MIX-SC	69.3	73.6	63.1	70.7
+ GraphOTTER	49.1	53.9	48.9	55.5
GPT-o4-mini (reference)	71.0	78.0	69.3	78.7

VII. PROPOSED CONCEPTUAL ARCHITECTURE

Fig. 1 presents a proposed conceptual architecture derived from the findings above; it is conceptual and has not been

experimentally validated. Scientific sources are processed by an extraction layer for entities, relations and table structure, including merged and multi-level headers (layers 1–2) [5], [6]; this addresses the drop in accuracy on hierarchical headers [4]. Extracted content is kept in a structured store with captions and source identifiers (layer 3).

A cell-selection layer chooses only the relevant rows, columns and metrics (layer 4), motivated by the benefit of selective cell use seen with GraphOTTER on large tables and by the metric-selection errors reported in [4].

A reasoning layer then works in explicit steps, which raised accuracy in [4], and delegates arithmetic to executable code (layer 5); because PoT alone scored below CoT in [4], code is used here only as a check, not as the sole method.

A verification layer recomputes values, cross-checks the chosen metric and flags uncertainty (layer 6).

Results pass to a researcher for review before they are used for hypotheses or reports (layer 7), and corrections feedback to extraction and selection.

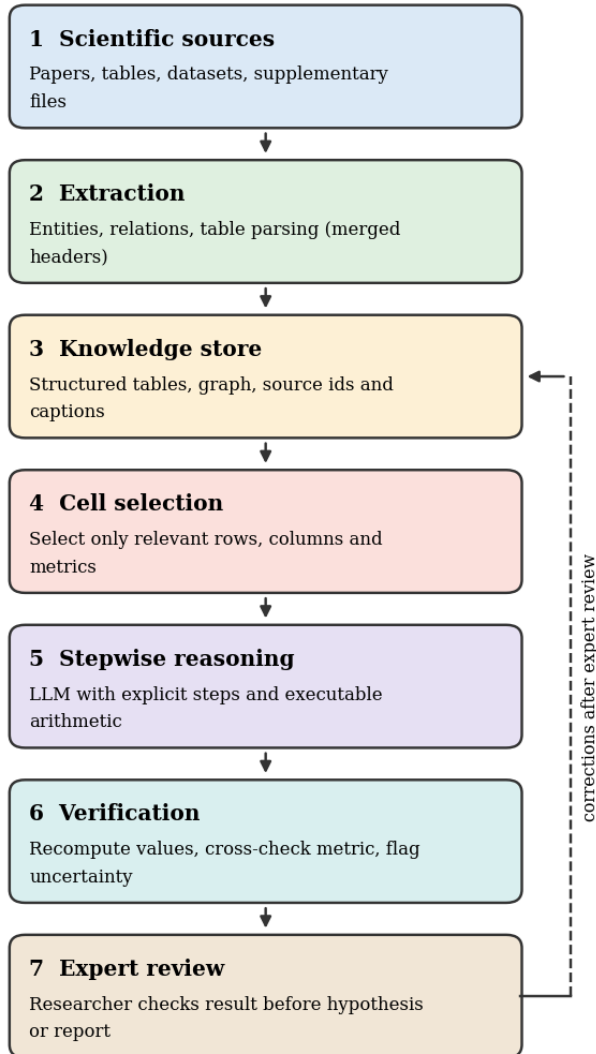


Fig. 1. Proposed conceptual architecture for AI-assisted analysis of scientific tables and literature (conceptual, not experimentally validated).

VIII. RESEARCH GAPS AND FUTURE DIRECTIONS

Gaps supported by the reviewed evidence. (i) Domain concentration: SciNumQA is centred on computer science, mainly computation and language and machine learning [4]. (ii) Generation and filtering bias: questions are produced and filtered by LLMs, which may over-represent certain patterns and favour items that are easy for the filter to judge [4]. (iii) Moderate human agreement: Fleiss' kappa was 0.55 on the

checked subset [4]. (iv) Low accuracy without explicit reasoning, with strong sensitivity to table size and hierarchical headers [4]. (v) Selection errors: wrong value or wrong metric even when all data were present [4]. (vi) Narrow task: the benchmark covers short-form question answering over standalone tables, not full discovery workflows [4]. (vii) Limited coverage of this review: it relies mainly on one benchmark paper and does not survey discovery applications such as [2], [3] in depth.

Directions requiring investigation. Priorities include benchmarks for other scientific domains such as biomedicine and chemistry, which the authors of [4] suggest can be reached by extending their pipeline; methods that handle hierarchical headers and large tables; combined reasoning over tables and surrounding text, as in SCITAT [10]; verification layers that recompute and cross-check results; evaluation of the reliability of LLM judges against human experts; and workflows in which researchers review outputs before they enter hypotheses.

IX. CONCLUSION

AI for scientific discovery spans extraction from text, question answering and, increasingly, reasoning over the tables that carry experimental evidence. Benchmarks have progressed from lookup-oriented tasks to multi-step numerical reasoning, and explicit reasoning prompts raise accuracy considerably, yet performance still falls on larger and hierarchically structured tables and errors in value and metric selection persist. The evidence base reviewed here is narrow, computer-science centred and partly dependent on LLM-generated data. A pipeline with cell selection, stepwise reasoning, verification and expert review, as outlined in the proposed architecture, is a reasonable direction but requires empirical validation across scientific domains.

REFERENCES

1. C. K. Reddy and P. Shojaee, "Towards scientific discovery with generative AI: Progress, opportunities, and challenges," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 27, 2025, pp. 28601–28609.
2. Y. Zheng, H. Y. Koh, J. Ju, A. T. N. Nguyen, L. T. May, G. I. Webb, and S. Pan, "Large language models for scientific discovery in molecular property prediction," Nature Mach. Intell., vol. 7, no. 3, pp. 437–447, Feb. 2025, doi: 10.1038/s42256-025-00994-z.
3. J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, pp. 583–589, 2021.

4. S. Son, Y. Kim, G. Kim, D. Choi, and M. Sung, "SciNumQA: A benchmark for complex numerical reasoning over scientific tables," *IEEE Access*, vol. 14, pp. 82812–82822, 2026, doi: 10.1109/ACCESS.2026.3697546.
5. Q. Zhang, Z. Chen, H. Pan, C. Caragea, L. J. Latecki, and E. Dragut, "SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2024, pp. 13083–13100.
6. Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, Oct. 2018, pp. 3219–3232.
7. X. Wang et al., "SciBench: Evaluating college-level scientific problem-solving abilities of large language models," in *Proc. 41st Int. Conf. Mach. Learn.*, 2023, pp. 50622–50649.
8. W. Xu, Y. Deng, H. Zhang, D. Cai, and W. Lam, "Exploiting reasoning chains for multi-hop science question answering," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Nov. 2021, pp. 1143–1156.
9. A. Ghosh, V. S. Bathini, N. Ganguly, P. Goyal, and M. Singh, "How robust are the QA models for hybrid scientific tabular data? A study using customized dataset," in *Proc. Lang. Resour. Eval. Conf.*, May 2024, pp. 8258–8264.
10. X. Zhang, D. Wang, B. Wang, L. Dou, X. Lu, K. Xu, D. Wu, and Q. Zhu, "SCITAT: A question answering benchmark for scientific tables and text covering diverse reasoning types," in *Proc. Findings Assoc. Comput. Linguistics: ACL*, Jul. 2025, pp. 3859–3881.
11. P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, Beijing, China, Jul. 2015, pp. 1470–1480.
12. Y. Zhao et al., "DocMath-Eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, Bangkok, Thailand, Aug. 2024, pp. 16103–16120.
13. N. S. Moosavi, A. Rücklé, D. Roth, and I. Gurevych, "SciGen: A dataset for reasoning-aware text generation from scientific tables," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, 2021, pp. 1–13.
14. M. Akhtar, A. Shankarampeta, V. Gupta, A. Patil, O. Cocarascu, and E. Simperl, "Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Singapore, Dec. 2023, pp. 15391–15405.
15. J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971, doi: 10.1037/h0031619.
16. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.
17. W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *Trans. Mach. Learn. Res.*, Oct. 2023, pp. 1–20.
18. T. Liu, F. Wang, and M. Chen, "Rethinking tabular data understanding with large language models," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, Mexico City, Mexico, Jun. 2024, pp. 450–482.
19. Q. Li, C. Huang, S. Li, Y. Xiang, D. Xiong, and W. Lei, "GraphOTTER: Evolving LLM-based graph reasoning for complex table question answering," in *Proc. 31st Int. Conf. Comput. Linguistics*, 2025, pp. 5486–5506.