Superhuman Synthesis of Scientific Knowledge with LLM Agents

Xueqing Xu
Department of Physics
University of Cambridge
Cambridge, UK
Email: xx823@cam.ac.uk
Word Count: 927

***Abstract*—This executive summary presents key findings from a comprehensive study that reproduced and extended the PaperQA2 framework for scientific literature synthesis. This astronomical extension work has been accepted at the ICML 2025 Workshop on Machine Learning for Astrophysics. Our cross-domain evaluation reveals that different scientific fields require fundamentally different RAG architectures, challenging universal AI approaches and providing actionable guidance for scientific researchers implementing AI-assisted research systems.**

I. Project Overview

Recent advances in large language models (LLMs) have shown remarkable promise for automating scientific literature synthesis [1], yet concerns about hallucination [2] and knowledge cut-off [3] have limited their adoption in critical research applications. While the groundbreaking PaperQA2 [4] system demonstrated superhuman performance on biology literature tasks, a fundamental question remained unanswered: can these achievements generalize across different scientific domains, or do different fields require fundamentally different approaches?

This study addresses this critical gap by reproducing and extending PaperQA2's framework to systematically evaluate retrieval-augmented generation (RAG) systems across scientific domains. Our work provides comprehensive analysis of how domain-specific characteristics influence AI system performance in scientific knowledge synthesis, with profound implications for autonomous scientific discovery applications. This work has been accepted for presentation at the ICML 2025 Workshop on Machine Learning for Astrophysics, demonstrating its significance to the scientific AI community.

II. Methodology

Our reproduction leveraged the open-source PaperQA2 package from FutureHouse Inc., configured with GPT-4o-mini/GPT-4o/GPT-4.1 as the primary language model across all components and text-embedding-3-small for document vectorization. Key operational parameters included evidence_k=30 for retrieval depth, answer_max_sources=15 for final answer generation, and 5000-character chunk sizes with 250-character overlap for optimal document segmentation.

We employed Inspect AI as the evaluation framework coordinating a multi-stage process where PaperQA2 generates evidence-based responses subsequently processed by AG2 agents for final answer extraction. Our systematic ablation studies removed key components (RCS, agentic architecture) to validate their individual contributions, while model comparison experiments tested GPT-4o-mini, GPT-4o, and GPT-4.1 across different system components to understand performance sensitivity to language model choice.

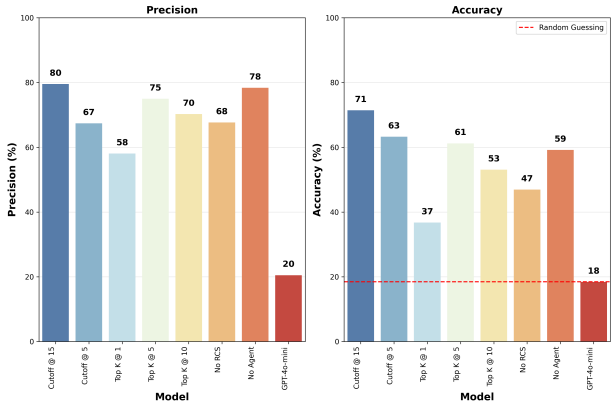

Fig. 1: PaperQA2 ablation study results on test dataset showing how removing key components (RCS, agentic architecture) dramatically impacts performance, with RCS being essential for biology but potentially problematic for astronomy applications.

To address the lack of standardized evaluation frameworks in astronomy, we developed SciRag, a comprehensive testing framework, and CosmoPaperQA2, a specialized benchmark containing 105 expert-curated cosmology questions from five influential papers spanning observational, theoretical, and computational aspects of modern cosmological research. We evaluated eight distinct RAG configurations (one without RAG for comparison) across multiple dimensions, including commercial systems (OpenAI Assistant, VertexAI, OpenAIPDF), hybrid approaches (HybridOAIGem, HybridGemGem), academic systems (PaperQA2 standard and modified), and baseline systems (Gemini, Perplexity). Our evaluation involved 945 expert assessments (9 systems × 105 questions) with domain

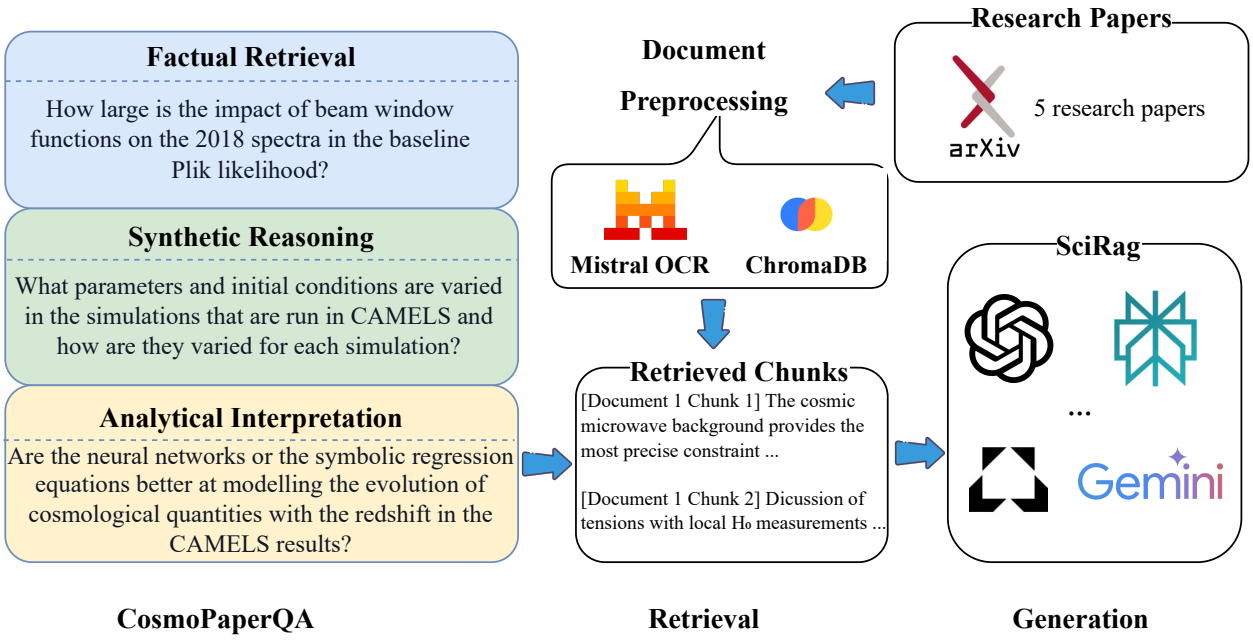

Fig. 2: SciRag System Architecture and CosmoPaperQA Benchmark Overview. Our framework integrates document preprocessing, retrieval mechanisms, and multi-provider generation to enable systematic evaluation of RAG Agents on astronomical literature.

experts holding PhD-level degrees in astronomy, astrophysics, or physics, ensuring authentic research standards.

III. Critical Findings

We successfully reproduced PaperQA2's performance on the LitQA2 biology benchmark, achieving 71% accuracy and validating the critical importance of both the reranking and contextual summarization (RCS) component and agentic architecture. Our systematic ablation studies confirmed that removing the RCS component dramatically reduces performance (from 80% to 68% precision), while eliminating the agentic architecture causes substantial accuracy drops (from 71% to 59% on test data).

Our most significant discovery challenges the prevailing assumption of universal RAG architectures. Commercial systems substantially outperformed PaperQA2 in astronomical tasks, with OpenAI Assistant achieving 91.4% accuracy, VertexAI reaching 86.7%, while PaperQA2 achieved 81.9% . This performance gap represents a fundamental shift in performance hierarchy compared to biology literature tasks.

The RCS component exemplifies domain sensitivity in RAG architecture. While RCS proved essential for biology literature tasks—improving precision from 68% to 80%—our astronomy evaluation suggests it may actually be detrimental for cosmological literature synthesis. The RCS step appears to remove or compress critical astronomical details, mathematical relationships, and technical specifics essential for accurate cosmology question answering.

IV. Cost-Performance Analysis

Our comprehensive cost-performance analysis reveals significant variations in efficiency across implementations with direct implications. VertexAI emerges as optimal for cost-conscious applications (86.7% accuracy with 13.3× cost advantage over GPT-4.1), while OpenAI systems achieve highest absolute performance (89.5-91.4%) but at premium costs. Hybrid approaches offer compelling balanced trade-offs (84.8-85.7% accuracy at moderate costs). These findings provide actionable guidance for institutions choosing optimal AI configurations based on their specific performance requirements and budget constraints.

V. Future Directions and Impact

Our findings reveal critical limitations that define the research frontier for scientific AI systems. While our evaluation spans biology and astronomy, systematic assessment across diverse scientific fields—chemistry, physics, materials science—is essential to establish comprehensive domain-specific optimization principles. Current evaluation frameworks remain constrained, with LitQA2 limited to multiple-choice biology questions and our CosmoPaperQA2 addressing only astronomical literature. Future work requires developing sophisticated evaluation metrics beyond accuracy and precision, incorporating

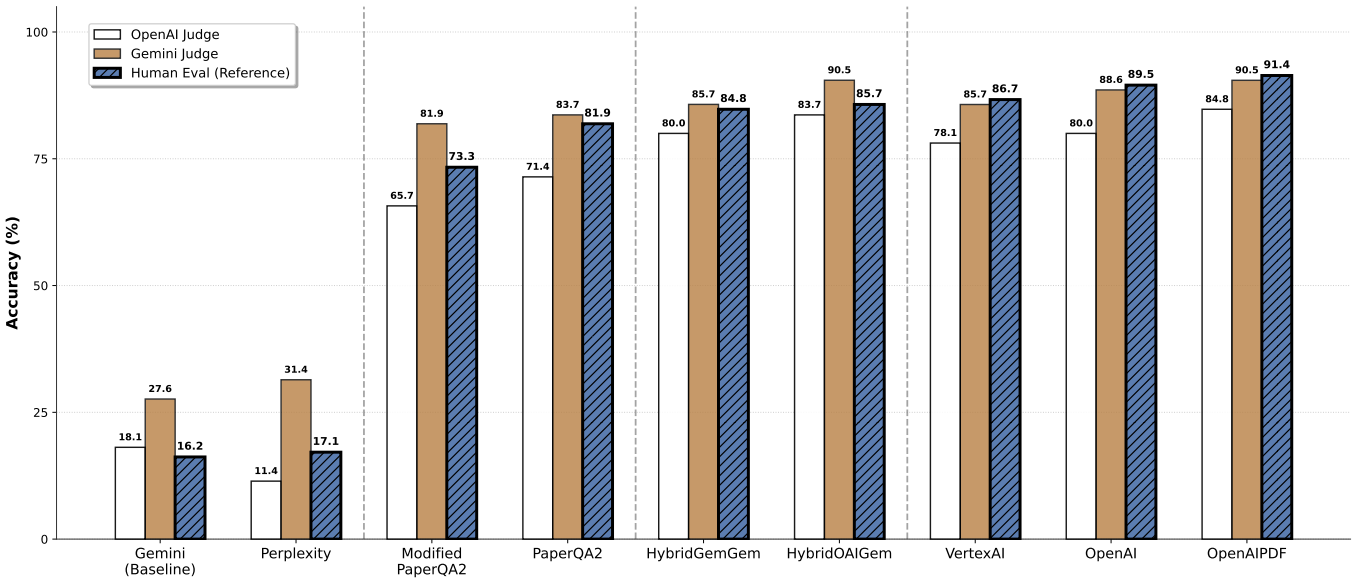

Fig. 3: Performance comparison of SciRag Agents across three evaluation methods showing four distinct performance tiers. Vertical dashed lines separate the four performance groups. The first two entries (Gemini Baseline and Perplexity) do not perform RAG but simply rely on pre-trained LLM knowledge and, for Perplexity, built-in retrieval tools.

reasoning quality assessment, source integration evaluation, uncertainty quantification, and response reliability measures critical for autonomous scientific discovery applications.

The computational limitations of our reproduction, including the unavailable Citation Traversal tool and single-run experiments, highlight the need for complete replication studies with full statistical comparison capabilities. More fundamentally, our work demonstrates that RAG agents have matured sufficiently to achieve expert-level performance on domain-specific scientific tasks, positioning AI systems as genuine research partners rather than information retrieval tools.

For practical deployment, researchers should prioritize domain-specific RAG optimization over generic solutions, requiring investment in specialized research computing infrastructure and domain expertise during system design. Our calibrated AI evaluators enable scalable assessment of thousands of scientific queries, essential for autonomous research applications, while our cost-performance framework provides immediate guidance for institutions balancing performance requirements with budget constraints—VertexAI for cost-conscious applications, OpenAI systems for peak performance needs, and hybrid implementations for competitive performance at reduced costs.

VI. Conclusion

This work successfully validates that RAG agents can achieve superhuman performance on scientific literature tasks while revealing the critical importance of domain-specific optimization. Our cross-domain evaluation demonstrates that the path to effective scientific AI lies not in universal solutions, but in carefully tailored architectures that respect the unique characteristics and requirements of different scientific disciplines. The practical implications extend far beyond academic interest to real-world deployment considerations, providing researchers with actionable guidance for implementing AI systems that can genuinely accelerate scientific discovery.

References

[1] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha, *The ai scientist: Towards fully automated open-ended scientific discovery*, 2024. arXiv: 2408.06292 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2408.06292.

[2] L. Huang, W. Yu, W. Ma, *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, Jan. 2025, ISSN: 1558-2868. DOI: 10.1145/3703155. [Online]. Available: http://dx.doi.org/10.1145/3703155.

[3] J. Cheng, M. Marone, O. Weller, D. Lawrie, D. Khashabi, and B. V. Durme, *Dated data: Tracing knowledge cutoffs in large language models*, 2024. arXiv: 2403.12958 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.12958.

[4] M. D. Skarlinski, S. Cox, J. M. Laurent, *et al.*, "Language agents achieve superhuman synthesis of scientific knowledge," *arXiv preprent arXiv:2409.13740*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2409.13740.

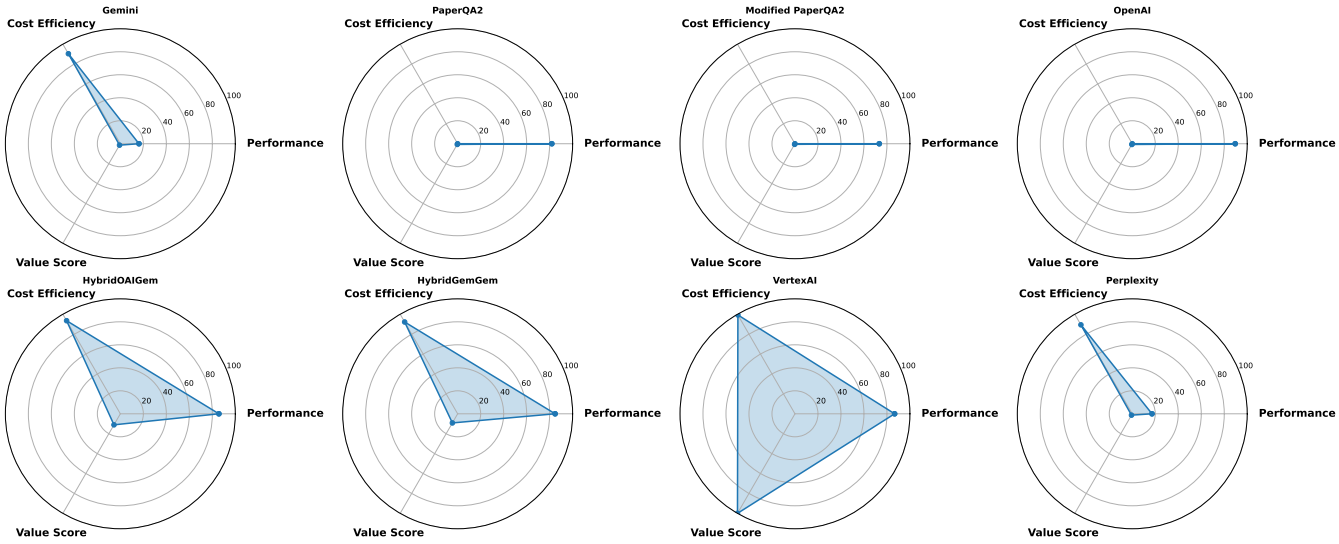

Fig. 4: Multi-dimensional performance analysis showing performance, cost efficiency, and value scores across RAG systems, with VertexAI demonstrating exceptional cost efficiency while OpenAI systems achieve peak performance at premium costs.