7: Superhuman Synthesis of Scientific Knowledge with LLM Agents

Abstract

Recent advances in large language models (LLMs) have shown promise for automating scientific literature synthesis, yet concerns about hallucination and domain-specific performance have limited their adoption in critical research applications. This study reproduces and extends the PaperQA2 framework, which demonstrated superhuman performance on biology literature tasks, by evaluating its effectiveness across scientific domains and developing domain-specific optimization strategies. We successfully reproduced PaperQA2's performance on the LitQA2 benchmark, achieving 71% accuracy on biology literature tasks and validating the critical importance of both reranking and contextual summarization (RCS) components and agentic architecture. 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. Our systematic evaluation of eight distinct RAG configurations (one without RAG for comparison) revealed significant domain-specific performance variations, with commercial systems (OpenAI Assistant: 91.4%, VertexAI: 86.7%) substantially outperforming PaperQA2 (81.9%) on astronomical tasks. Critically, we discovered that the RCS component, essential for biology literature synthesis, may be detrimental for cosmological research due to information dilution effects that remove crucial mathematical relationships and technical details. Our cost-performance analysis identified VertexAI as optimal for resource-conscious applications, while hybrid approaches achieved competitive performance at reduced costs. These findings fundamentally challenge one-size-fits-all RAG approaches and demonstrate that different scientific domains require tailored architectures, with domain-specific optimization being essential rather than merely beneficial for achieving expert-level performance in scientific knowledge synthesis.

Content

Contents

Word Count: 6811

AI Declaration

GitHub Copilot was used during development to assist with code completion, function documentation, debugging, and refactoring suggestions. All AI-generated code underwent thorough review to ensure correctness, adherence to project requirements, and proper error handling. The core algorithms and methodological decisions remain my intellectual contribution, with Copilot serving only as a productivity tool for implementing standard techniques and reducing time spent on repetitive coding tasks.

1 Introduction

Recent advances in large language models (LLMs) have shown promise in automating various aspects of scientific search, yet concerns about hallucination and factual accuracy have limited their adoption in critical scientific research. The work done by [1] Skarlinski et al. challenged this by demonstrating their designed PaperQA2 RAG systems not only match but exceed human experts' work on the LitQA2 benchmark, designed to synthesize the real research scenarios in biology. PaperQA2 achieved superhuman precision on literature research tasks and did not differ significantly from humans in accuracy.

Despite these successes in biology, astronomy presents distinct methodological challenges that require domain-specific RAG evaluation. As annotated by Bowman et al. [2], developing human-annotated benchmarks for doctoral-level scientific research domains remains economically prohibitive. Consequently, systematic evaluation of RAG Agents in astronomy is limited by the lack of standardized benchmarks and authentic evaluation datasets that capture the complexity of real research scenarios.

To address this gap, we developed SciRag, a comprehensive testing framework specifically for astronomy research applications. The framework includes CosmoPaperQA, a specialized benchmark containing 105 expert-designed questions and answers drawn from five influential cosmology papers. Using this system, we can systematically evaluate and compare different types of AI research assistants - from commercial platforms like OpenAI to specialized tools like PaperQA2 and search-enhanced platforms like Perplexity. The ultimate aim is to provide clear, practical recommendations for scientists choosing the most effective AI configuration for their astronomy research needs.

1.1 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) has become a widely adopted technique for enhancing the reliability of Large Language Models through the incorporation of external contextual information in their responses. RAG enables LLMs to address several challenges: maintaining current information beyond training[3], preventing leakage of sensitive private data[4], and most importantly, reducing false information generation while improving factual precision within a specific subject domain [5].

Traditional RAG systems operate through a multi-stage pipeline that begins with document ingestion and preprocessing. During the ingestion phase, raw documents undergo chunking strategies that segment content into retrievable units, typically ranging from 100 to 1000 tokens depending on the application domain[6]. These chunks are then converted into dense vector representations using embedding models such as sentence transformers or domain-specific encoders. The resulting embeddings are stored in vector databases optimized for similarity search, such as FAISS, Pinecone, or Chromadb[6].

During the retrieval phase, user queries undergo the same embedding transformation, enabling semantic similarity matching against the stored knowledge base. Multiple RAG frameworks have been designed to enhance retrieval methodologies, including semantic-based search[7], hybrid search approaches[8], integrated search with re-ranking mechanisms[9], and context-aware retrieval systems[10]. The retrieved contexts are then ranked by relevance scores, with top-k passages selected for inclusion in the generation prompt. Finally, the LLM synthesizes responses by conditioning on both the original query and the retrieved contextual information, ideally producing more accurate and grounded outputs[11].

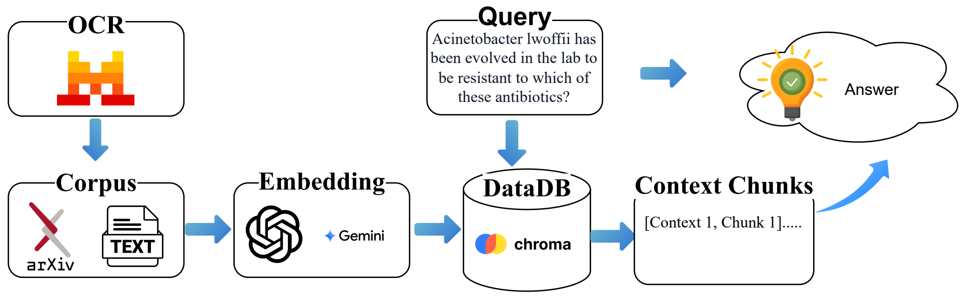

Figure 1: RAG workflow.

RAG systems show considerable promise for improving large language model capabilities, but their success depends critically on the quality and organization of the underlying data. Knowledge databases are often built by using OCR technology to convert unstructured information from PDF files into structured formats. Unfortunately, OCR technology isn't perfect—its prediction errors become embedded within the knowledge base, which compromises the accuracy of the RAG system's outputs. To address this challenge, we will use Mistral OCR [1], a more advanced solution.

1.2 Applications in Academia

Incorporating LLM and RAG into the academic domain promises to enable a more efficient processing of scientific literature, assisting scientists in new discoveries. Yet academic documents, with their sophisticated layouts and mathematical formulas, create implementation challenges that differ significantly from typical conversational RAG scenarios. The following are some RAG assistants used in academia:

1. **PaperQA2:** PaperQA2 [1] employs a multi-agent architecture that decomposes RAG into specialized tools, allowing the system to iteratively refine its search and reasoning process rather than using a fixed pipeline. The core agent orchestrates four main tools: a **Paper Search** tool that transforms user queries into keyword searches and parses documents, a **Gather Evidence** tool that performs top-k vector retrieval followed by LLM-based reranking and contextual summarization (RCS), a **Citation Traversal** tool that exploits citation graphs to find additional relevant sources, and a **Generate Answer** tool that synthesizes final responses from ranked evidence summaries.

2. **OpenScholar:** The system [12] retrieves relevant content from a database of 45 million open-access academic papers. It employs a cyclical self-improvement framework in which a language model first produces a preliminary answer, then evaluates its own work using natural language critiques (such as noting gaps in information or structural problems), and subsequently leverages these critiques to conduct further searches and

[1] https://mistral.ai/news/mistral-ocr

enhance the final response. The research demonstrates that OpenScholar-8B achieves superior accuracy compared to PaperQA2 when evaluated on the ScholarQABench dataset.

In astronomy, Ciucă et al. [13] showed by in-context learning and adversarial prompting, LLMs can synthesize diverse astronomical information into coherent and innovative hypotheses, establishing that astronomical applications could benefit significantly from document retrieval capabilities. Similarly, Shao et al. [14] demonstrated that LLMs can effectively extract specialized astronomical knowledge entities from astrophysics journals using carefully designed prompting strategies. Recent work by Wu et al. [15] has realized the critical need for systematic evaluation of RAG Agents in astronomical research. They proposed a dynamic evaluation framework using a Slack-based chatbot that retrieves information from arXiv astro-ph papers, emphasizing the importance of real-world user interactions over static benchmarks. While their approach provides valuable insights into user behavior and system usability, it relies on user feedback and reaction data rather than systematic performance assessment against validated ground-truth. This highlights a complementary need for standardized benchmarks that can provide consistent, reproducible evaluation metrics across different RAG implementations.

1.3 Benchmarks and Evaluation in Cosmology

Existing evaluation falls into two categories, each with some limitations:

Astronomy-Specific Knowledge Benchmarks: AstroMLab 1 [16] provides the first comprehensive astronomy-specific evaluation with 4425 AI-generated multiple-choice questions from Annual Review articles. While demonstrating significant performance variations between models with specialized astronomical knowledge, its multiple-choice format and automated question generation limit evaluation to content mastery rather than scientific inquiry workflows. Similarly, Astro-QA [17] provides a structured evaluation with 3082 questions spanning diverse astronomical topics. However, its synthetic questions limit its ability to assess the complex, open-ended reasoning required for an authentic scientific research workflow.

General Scientific Evaluation: Broader scientific benchmarks like LitQA2 [1], ChemRAG-Toolkit [18], ScisummNet [19] are designed for other scientific domains and may not capture astronomy-specific challenges such as mathematical reasoning about cosmological models, and interpretation of observational constraints.

1.4 Reproduction Goals

This reproduction study seeks to verify the performance of PaperQA2 [1] by utilizing the publicly available framework to replicate critical experimental findings depicted in Figure 2. While we use the open-source PaperQA2 implementation, the Citation Traversal tool, which remains unavailable to the public, constrains our reproduction scope to Figure 2 panels A, B, and C, excluding the DOI recall analysis in panel D.

The main reproduction focuses on validating whether the reported superhuman performance can be independently achieved using identical language models evaluated on the LitQA2 benchmark. Two subsequent research directions complement this core investigation: First, we examine the relative importance of individual PaperQA2 components in delivering performance advantages over human experts through systematic ablation studies. Second, we investigate how variations in model architecture and hyperparameter configurations influence the system's capacity for accurate scientific knowledge synthesis, thereby testing the robustness of the original paper's claims regarding optimal system design.

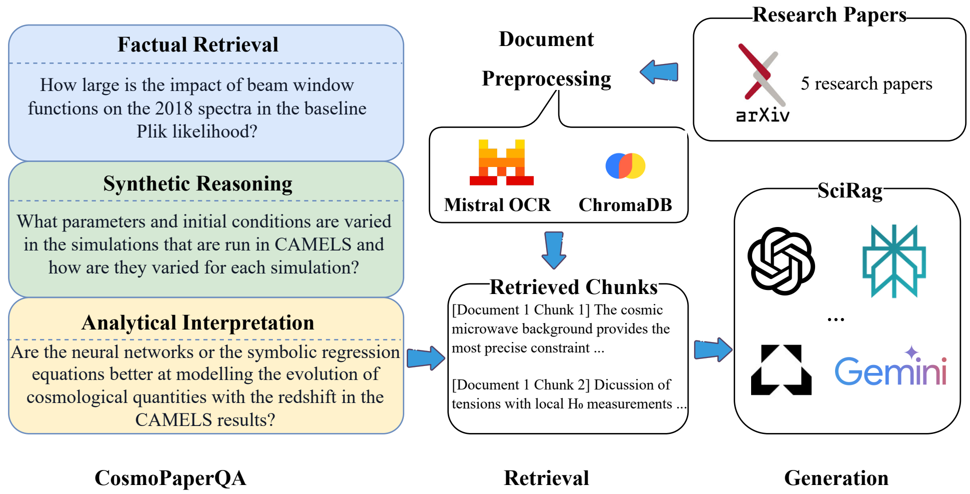

Figure 2: SciRag System Architecture and CosmoPaperQA Benchmark Overview. Our framework integrates document preprocessing, retrieval mechanisms, and multi-provider generation to enable systematic evaluation of RAG Agents on astronomical literature.

1.5 Extension

Building on this reproduction work, we extend the evaluation framework to astronomical research through a comprehensive assessments of RAG agents performance in cosmology. We evaluated eight distinct RAG agents (one without RAG for comparison) configurations using 105 expert-curated cosmology question-answer pairs developed specifically for this study. Each configuration was evaluated by a human expert, resulting in an assessment of 945 generated responses.

Our results demonstrate that the optimal RAG configuration, OpenAI Embedding, achieves 91.4% accuracy. Using these human evaluation results, we calibrate AI evaluators that serve as reliable proxies for human assessment. This approach enables systematic selection of optimal RAG configurations for multi-agent systems in autonomous astrophysical discovery while providing scalable AI evaluators capable of handling thousands of cosmology question-answer pairs.

2 Methodology

2.1 Dataset Preparation

To evaluate the performance, two evaluation benchmarks were used.

2.1.1 LitQA2

LitQA2 is a set of 248 multiple-choice questions (49 in test dataset and 199 in training dataset) with answers that require retrieval from scientific literature, designed to have answers that appear in the main body of a paper, but not in the abstract. The authors specify that *question authors were instructed to identify recent papers (published within the last 36 months)* to ensure the questions test the system's ability to retrieve current scientific knowledge.

Our detailed examination uncovers notable structural differences across the dataset. Questions within both training and test sets feature 3-12 response options (mean: 5.39 options per question in training, 5.61 options per question in testing). When accounting for these varying choice distributions, the expected random performance baseline stands at 22.4% for the training set and 18.4% for the test set.

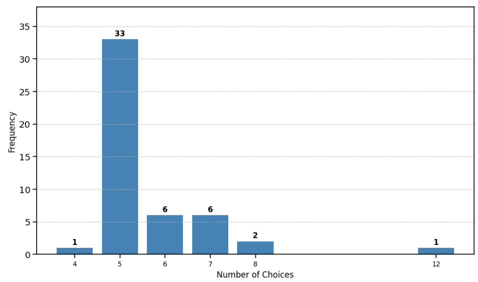

Figure 3: Distribution of Number of Choices per Question in Test Dataset.

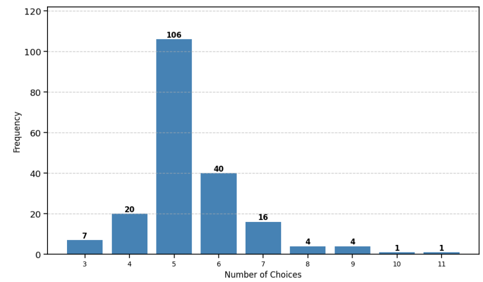

Figure 4: Distribution of Number of Choices per Question in Train Dataset.

The choice randomization mechanism operates through the `randomize_choices()` function, which combines the correct answer, the fixed abstention option (*Insufficient information to answer the question*), and variable distractor sets into a unified choice pool. The system implements shuffling to prevent positional bias, dynamically assigns alphabetical labels (A through L) based on choice count, and tracks both the correct answer position and abstention option location for evaluation purposes. Output formats include JSON, JSONL, and CSV to support different evaluation frameworks.

However, our analysis of the dataset composition reveals a significant discrepancy between the stated methodology and the actual temporal distribution of source papers. While the test dataset (49 questions) adheres strictly to the recency criteria with all questions derived from papers published in 2024, the training dataset (199 questions) spans a much broader temporal range. Our examination shows that papers in the training set date back to 2000, with the distribution heavily skewed toward recent years but including papers that are over two decades old, clearly violating the stated 36-month recency requirement.

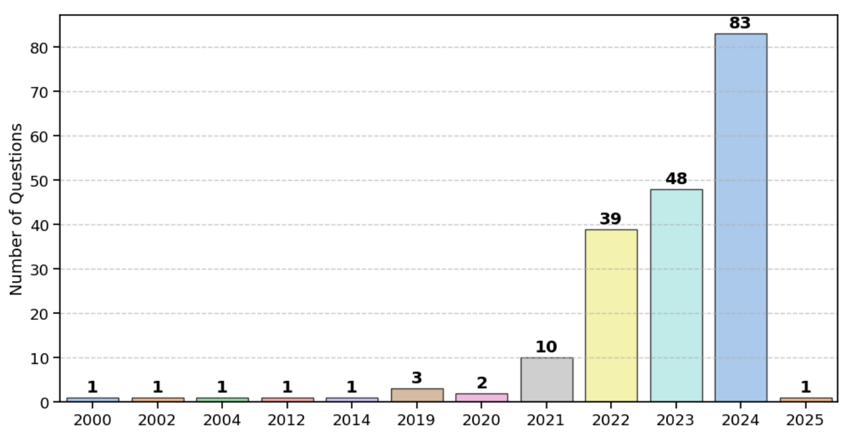

Figure 5: Distribution of LitQA2 Training Questions by Year.

Furthermore, the question generation strategy differs between the training and test sets. The test dataset follows a one-question-per-paper approach, with all 49 questions generated from unique 2024 publications. In contrast, the training dataset allows multiple questions from the same paper, with 174 papers contributing 1 question each, 8 papers contributing 2 questions each, and 3 papers contributing 3 questions each.

This discrepancy suggests either that the stated guidelines were not strictly enforced during dataset creation, or that exceptions were made to the recency criteria that were not adequately documented in the paper's methodology section. This finding has important implications for interpreting benchmark results, as the training set's inclusion of older papers may affect the generalizability of models trained on this data.

Given these findings, we analyze the training and test datasets separately throughout our evaluation of PaperQA2's performance, placing greater trust in the test dataset results as they better reflect the original paper's stated methodology and provide a more reliable benchmark for assessing performance on current scientific literature.

2.1.2 CosmoPaperQA2

To address the evaluation challenges in astronomy-specific RAG assessment, Adrian and Boris developed CosmoPaperQA2 [20], a specialized benchmark for cosmological research applications. Unlike existing general-purpose scientific benchmarks, this dataset captures the unique methodological challenges inherent to astronomical research, including mathematical reasoning about cosmological models and interpretation of observational constraints.

Questions within the benchmark originate from five highly influential papers spanning critical areas of modern cosmology: the Planck 2018 cosmological parameters [21], CAMELS machine learning simulations [22], [23], local Hubble constant measurements [24], and recent Atacama Cosmology Telescope constraints [25]. This curation ensures comprehensive coverage of observational, theoretical, and computational aspects of modern cosmological research.

CosmoPaperQA2 is specifically designed to evaluate three key capabilities essential for astronomical research: **zero-shot learning performance** on previously unseen question types, **open-ended question answering** that mirrors real research scenarios rather than constrained multiple-choice formats, and **multi-source knowledge synthesis** requiring integration across observational, theoretical, and computational domains. This design enables comprehensive assessment of our RAG systems' ability to support genuine astronomical research workflows.

2.2 PaperQA2 Implementation Details

2.2.1 System Architecture and Configuration

The implementation leveraged the open source PaperQA2 package developed by FutureHouse Inc.[2], building upon their agentic RAG framework rather than developing a custom solution from scratch. The system was configured with GPT-4o-mini/GPT-4o/GPT 4.1 as the primary language model across all components (question answering, summarization, and agent operations) with temperature settings of 0.1 for deterministic responses and rate limiting at 30,000 requests per minute. Text-embedding-3-small was used for document vectorization and retrieval operations, creating hybrid dense-sparse embeddings for enhanced semantic retrieval.

Key operational parameters were established as baseline configurations (these parameters are systematically varied in subsequent experiments):

[2] Available at: https://github.com/Future-House/paper-qa

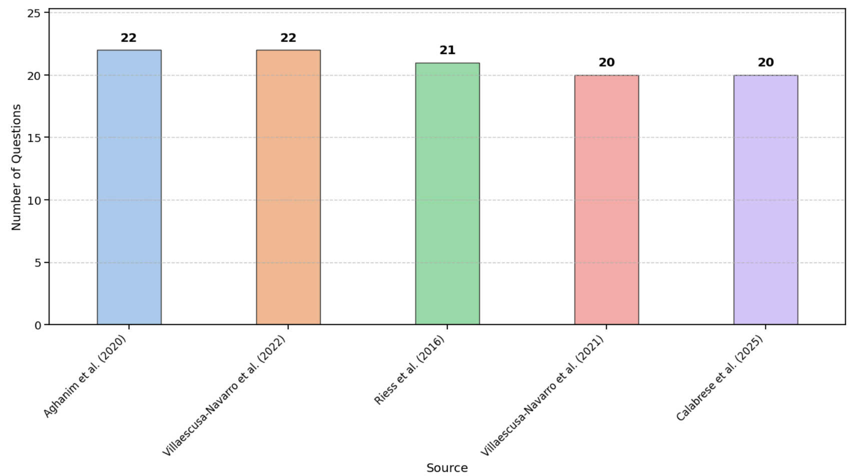

Figure 6: Distribution of CosmoPaperQA Questions by Source.

- **evidence_k**: This parameter was set to 30 by default in LitQA experiments, and Top-k Rank @ X experiment where it was set to X.
- **answer_max_sources**: This parameter was 5 in the top-performing Answer cutoff @ 5, but 15 for all other experiments.
- **evidence_skip_summary**: This parameter was varied in the No RCS ablation experiments to evaluate the impact of the reranking and contextual summarization component.
- **chunk_size**: 5000 character segments with 250-character buffer zones. This parameter was varied in the experiments shown in Figure 13.
- **embedding**: text-embedding-3-small

2.2.2 Agentic Workflow Architecture

The study utilized PaperQA2's pre-built agentic architecture, which decomposes RAG into specialized tools, allowing the system to iteratively revise search parameters and examine candidate answers before producing final responses. The open source framework provided three primary tools orchestrated by an LLM-based agent:

1. **Paper Search Tool**: Transforms user requests into keyword searches to identify candidate papers, which are then parsed into machine-readable text and chunked for later usage
2. **Gather Evidence Tool**: Implements PaperQA2's signature two-phase process with initial embedding-based ranking followed by LLM-based reranking and contextual summarization (RCS)
3. **Generate Answer Tool**: Synthesizes top-ranked evidence summaries into final responses with proper citations

2.2.3 Document Processing and Parsing

The implementation utilized PaperQA2's integrated document processing capabilities, employing the built-in `get_directory_index()` function to recursively process the papers directory. We used PaperQA2's implementation of the Grobid document parsing algorithm, which reliably parses sections, tables, and citations from papers. The open source framework's automated metadata fetching capabilities retrieved information redundantly from multiple providers including Crossref and Semantic Scholar, incorporating citation counts and journal quality data without requiring custom API integration.

2.2.4 Question Processing Pipeline

Questions were processed sequentially with progress tracking and systematic JSON serialization of results, capturing question text, choices, PaperQA2 responses, correct answers, and abstention options. The Settings framework managed configuration parameters including model specifications, rate limiting, and directory paths for reproducible experiments. We employed function calling with MCAnswer schema for reliable letter extraction.

2.3 SciRag Implementation Details

2.3.1 Scientific Document Processing Pipeline

Our preprocessing pipeline addresses the requirements of astronomical literature through multi-stage processing. Optical character recognition (OCR) integration using Mistral's advanced capabilities handles tables, figures, mathematical expressions, and specialized notations common in astrophysics papers. This system generates multiple output formats, ensuring compatibility across different RAG backends. Our entire pipeline is available online.[3]

Document segmentation employs LangChain with 5000-token chunks and 250-token overlap, optimized for scientific text coherence. Special handling accommodates the 4096-token constraint of OpenAI Assistant while maintaining consistency across all implementations.

2.3.2 Multi-Backend Implementation For Comparative Analysis

All generation agents are configured with a temperature of 0.01 for consistent, deterministic responses, and top-k=20 (retrieving the 20 most similar document chunks per query) excluding Gemini Assistant, PaperQA2 (both versions) and Perplexity Assistant. The implementation provides both semantic search and hybrid retrieval capabilities across different backends, with specific configurations optimized for each system's strengths. Here are the configurations that we used for each assistant.

OpenAI Assistant: Direct implementation of OpenAI vector stores with file search tool (providing automatic query rewriting, parallel searches, keyword+semantic search, and result reranking) with text-embedding-3-large [26] for embeddings and GPT-4.1 for generation, with configurable retrieval parameters (similarity threshold=0.5).

OpenAIPDF Assistant: Direct PDF processing implementation without OCR preprocessing, enabling comparison of raw PDF handling versus OCR-enhanced document processing. Identical configuration to OpenAI Assistant, but operates on unprocessed PDF documents.

VertexAI Assistant: Google Cloud implementation using Google's text-embedding-005 for embeddings and gemini-2.5-flash-preview-05-20 [27] for generation. Creates RAG corpora through Vertex AI infrastructure with automatic document ingestion from Google Cloud Storage buckets. Supports semantic search with configurable similarity thresholds (0.5).

Gemini Assistant: Direct integration with Google's Gemini model gemini-2.5-flash-preview-05-20 for baseline comparison without specialized RAG infrastructure.

[3]https://github.com/cmbagent/scirag

HybridGemGem Assistant: Dual-Gemini implementation using Gemini's text-embedding-001 for embedding, leading embedding model on MTEB [28] [4]with ChromaDB storage and gemini-2.5-flash-preview-05-20 for generation. Supports ChromaDB backends with semantic-only search.

HybridOAIGem Assistant: Cross-platform architecture identical to HybridGemGem but specifically configured with OpenAI embeddings (text-embedding-3-large) and gemini-2.5-flash-preview-05-20, enabling comparison of embedding-generation combinations.

PaperQA2: Standard academic RAG implementation utilizing GPT-4.1 across all components (search, summarization, retrieval), evidence retrieval k=30, maximum 5 citations per response (optimal settings from original work). Processes OCR-enhanced documents with semantic-only search.

Modified PaperQA2: Domain-adapted version with identical technical configuration but specialized astronomical prompts and cosmological citation protocols. Uses evidence retrieval k=10 (reduced from standard k=30) for more focused responses.

Perplexity Assistant: Web-search enabled system using sonar-reasoning-pro model with real-time access to current literature. No local vector storage - relies entirely on web retrieval.

This diverse implementation suite enables comprehensive comparison across commercial, academic, and hybrid approaches, providing empirical guidance for selecting optimal RAG configurations for autonomous scientific discovery workflows. The detailed RAG Assisstants prompts are in Appendix A.

2.4 Evaluation Pipeline

Two distinct evaluation frameworks were built to analyze the multiple-choice and summary-based outputs generated by two benchmarks.

2.4.1 LitQA2

The LitQA2 evaluation pipeline uses Inspect AI [29] as the evaluation framework to coordinate a multi-stage evaluation process where PaperQA2 generates evidence-based responses that are subsequently processed by AG2 [30] agents for final answer extraction. The detailed AG2 prompt is in Appendix C. Inspect AI's @task decorator defines the evaluation parameters and manages the execution flow through its bridge pattern, which connects the custom PaperQA2 agent function to the framework's scoring and metrics system.

The evaluation process begins when Inspect AI loads questions from the LitQA2 dataset via `json_dataset()` and transforms them into standardized Sample objects containing the question text, multiple-choice options, and target answers. Within Inspect AI's execution framework, each sample is processed through the `paperqa2_agent()` function implementing a two-stage evaluation approach.

PaperQA2 first performs literature search using its RAG pipeline to generate comprehensive, cited responses from scientific papers' database. This evidence-based output is then passed to an AG2 researcher agent (GPT 4o-mini) that analyzes PaperQA2's response alongside the multiple-choice options, extracting the appropriate answer letter through logical analysis of the scientific content.

Inspect AI manages the evaluation workflow using the bridge() solver to connect the custom agent function with its scoring system. The `precision_choice()` scorer evaluates

[4]Retrieved on 30-05-2025

AG2's extracted answers against ground truth targets, implementing custom logic to handle *Insufficient information responses* with NOANSWER values. This enables calculation of three metrics: precision (fraction of questions answered correctly when a response is provided), accuracy (fraction of correct answers over all questions), and coverage (response rate), leveraging Inspect AI's task management capabilities to accommodate the specialized PaperQA2-to-AG2 evaluation pipeline.

2.4.2 CosmoPaperQA2

Domain experts are provided (1) a question query, (2) an ideal solution validated by experts, and (3) an RAG Agent-generated response. Then, evaluation is based on:

Correct (1): Generated responses demonstrate factual accuracy, and capture essential scientific understanding equivalent to the ideal answer.

Incorrect (0): Generated responses contain errors, contradict established scientific knowledge, or fail to include all the core concepts of ideal answers.

After obtaining the scores, we scaled them to 0-100 for comparison between different system configurations.

Cosmologists who evaluated the response are domain experts with PhD-level degrees currently working as researchers or faculty in astronomy, astrophysics, or physics. Together with these cosmologists, we designed the evaluation criteria and pipeline to ensure alignment with authentic research standards. In total, our experts evaluated 945 responses (9 systems $\times$ 105 questions) generated by RAG Agents.

We also explored LLM-as-a-Judge [31], an AI-based evaluation system calibrated for scientific research queries, using a binary scoring protocol aligned with human expert methodology. Our prompting experiments in Appendix B revealed that chain-of-thought, which asks models to formulate their underlying reasoning process, typically enhances evaluation accuracy and improves concordance with field expert judgments.

To investigate the bias of the pipeline specifically, as LLM evaluators may prefer responses generated by themselves [32], we used two LLM-as-a-Judge settings. Given that majority of generation systems utilize either OpenAI or Gemini-based agents, with the exception of the Perplexity Agent, we used the OpenAI o3 mini and Gemini gemini-2.5-pro-preview-06-05, reasoning models for evaluation.

3 Results

3.1 Evaluation Results on LitQA2

We conducted a comprehensive evaluation of our PaperQA2 implementation on the LitQA2 benchmark, systematically varying key parameters to understand their impact on performance. All experiments were conducted using GPT-4o-mini as the base model unless otherwise specified. Our evaluation focused on reproducing the core findings from the original PaperQA2 paper while investigating the sensitivity of the system to various configuration choices.

3.1.1 Overall Performance and Comparison with Original Work

Our reproduction of PaperQA2 on the LitQA2 benchmark demonstrates strong performance that aligns with the trends reported in the original work, though with some variations due to implementation differences and model updates. Figure 7 illustrates a representative example of how PaperQA2 processes a scientific question, retrieving relevant evidence and providing a well-reasoned answer with proper citations.

It is important to note that our evaluation represents single-run experiments for each configuration, unlike the original work which conducted 3 runs to compute statistical measures and confidence intervals. Additionally, rather than conducting human performance evaluation as in the original study, we calculated random guessing probability as our baseline comparison, which stands at 18.4% for the test set and 22.4% for the training set based on the varying number of multiple-choice options per question.

The baseline configuration (Cutoff @ 15) achieved 71% accuracy on test and 67% on training, while the optimal configuration reported in the paper (Cutoff @ 5) reached 63.0% on test and 67% on training accuracy. When comparing our results with the original PaperQA2 implementation, we observe that our reproduction successfully captures the key performance characteristics, with precision scores ranging from 20-87% and accuracy scores from 18-71% across different configurations and datasets.

The original work demonstrated PaperQA2 achieving superhuman precision while matching human accuracy, and our reproduction maintains this general performance characteristic, though with notable variations between the training and test datasets that provide additional insights into system robustness.

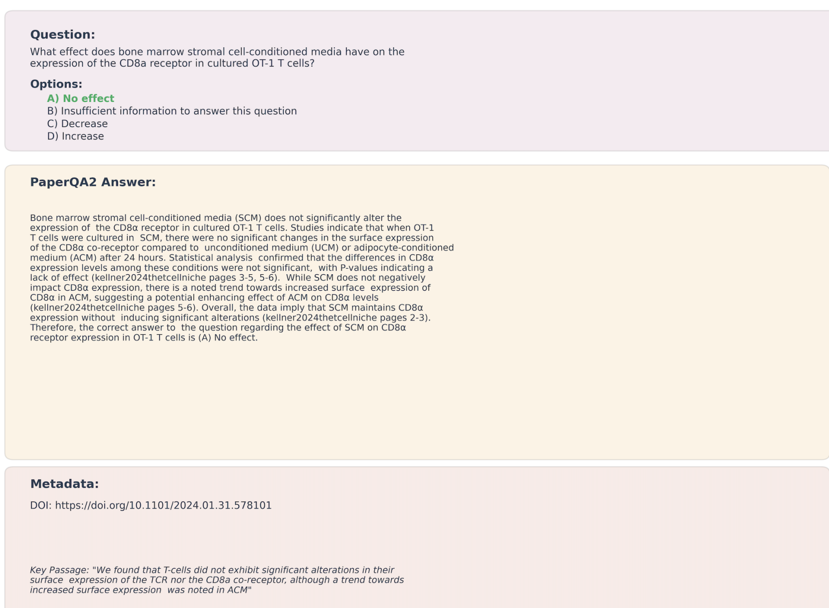

Figure 7: Representative example of PaperQA2 answering a LitQA2 question

3.1.2 Training vs. Test Dataset Performance Comparison

The systematic comparison between training and test datasets reveals performance differences that provide important insights about dataset composition and model capabilities.

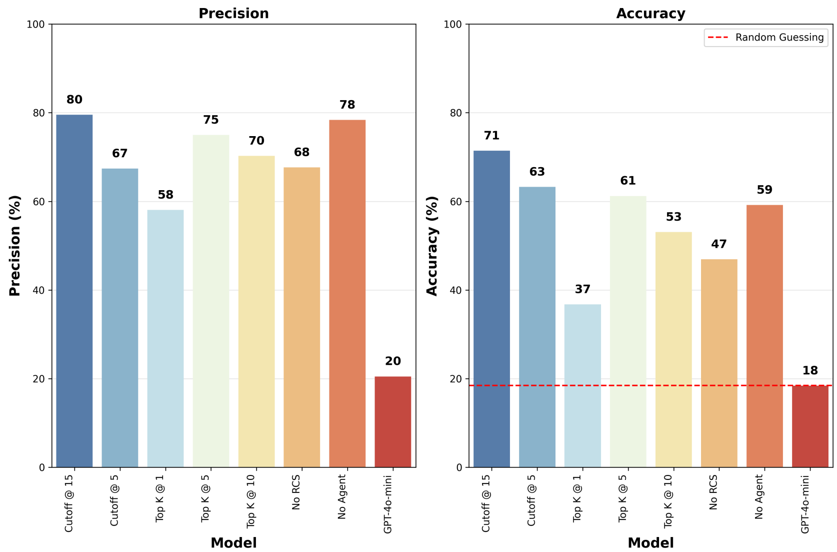

Figure 8: PaperQA2 performance study on LitQA2 test set. **Cutoff @ 15/5** refers to the max_sources parameter controlling the number of top-ranked evidence summaries included in final answer generation. **Top k @ 1/5/10** indicates the consider_sources parameter determining how many document chunks are processed in the RCS (Reranking and Contextual Summarization) step. **No RCS** represents ablation without the reranking and contextual summarization component. **No Agent** shows performance with a hard-coded sequential pipeline instead of PaperQA2's adaptive agentic architecture. **GPT-4o-mini** demonstrates performance with a language model only.

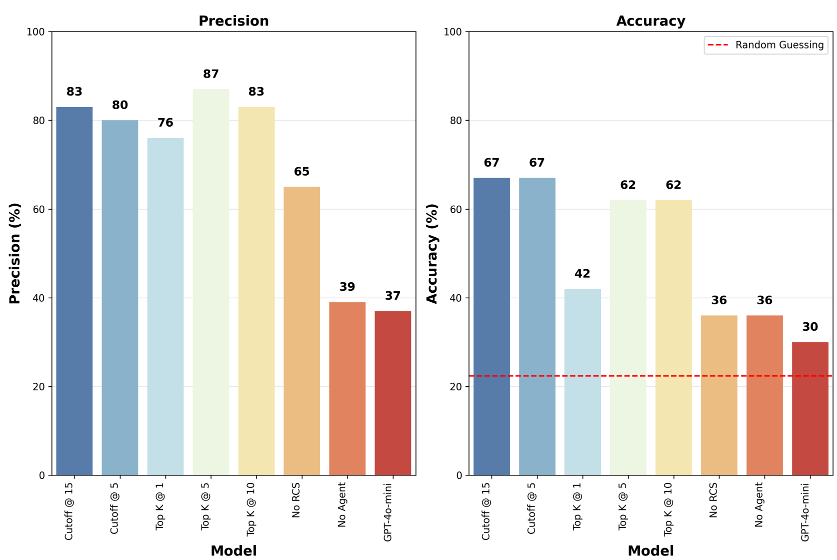

Figure 9: PaperQA2 performance study on LitQA2 train set. **Cutoff @ 15/5** refers to the max_sources parameter controlling the number of top-ranked evidence summaries included in final answer generation. **Top k @ 1/5/10** indicates the consider_sources parameter determining how many document chunks are processed in the RCS (Reranking and Contextual Summarization) step. **No RCS** represents ablation without the reranking and contextual summarization component. **No Agent** shows performance with a hard-coded sequential pipeline instead of PaperQA2's adaptive agentic architecture. **GPT-4o-mini** demonstrates performance with a language model only.

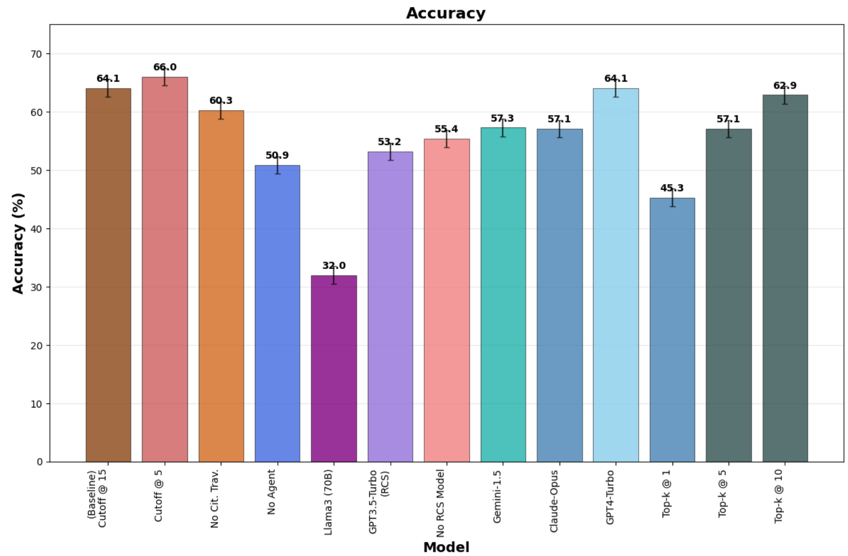

Figure 10: Original Work

Most notably, GPT-4o-mini (representing the base language model without PaperQA2 retrieval augmentation) exhibits a dramatic performance degradation from the training set (30% accuracy, 37 precision) to the test set (18% accuracy, 20% precision), representing a 40% relative decrease in accuracy and 46% decrease in precision.

This drop is particularly revealing given that the test dataset exclusively contains 2024 publications, which likely fall beyond GPT-4o-mini's training knowledge cutoff, forcing the model to rely solely on its pre-existing knowledge without access to the specific scientific content being queried, matching the percentage of random guessing. In contrast, the training dataset's broader temporal span (2000-2024) includes many papers that were likely present in GPT-4o-mini's training data, explaining the model's relatively better performance on this subset.

3.1.3 Core Component Ablation Analysis

The ablation analysis reveals the critical importance of PaperQA2's individual components through systematic removal experiments. The most severe performance degradation occurs when removing the RCS component (No RCS), which drops precision from 80% to 68% on the test set and from 83% to 65% on the training set, while accuracy falls dramatically to 47% (test) and 36% (training). This substantial impact demonstrates that the RCS step is essential for filtering irrelevant chunks and organizing retrieved evidence for this multiple-choice answer task.

Removing the agentic architecture (No Agent) also causes performance loss, reducing test set accuracy from 71% to 59% while maintaining relatively high precision at 78%. On the training set, the impact is more pronounced with accuracy dropping to 36%, highlighting that PaperQA2's iterative, adaptive search strategies are fundamentally superior to fixed pipeline processing for complex scientific literature tasks.

The evidence integration experiments reveal important optimization opportunities. Comparing cutoff strategies, Cutoff @ 5 achieves 67% precision and 63% accuracy on the test set, while Cutoff @ 15 reaches 80% precision and 71% accuracy, suggesting that including more evidence sources improves overall performance. The Top-k parameter analysis shows that moderate retrieval depth provides optimal balance: Top k @ 5 achieves 75% precision and 61% accuracy on the test set, outperforming both shallow retrieval (Top k @ 1: 58% precision, 37% accuracy) and deeper retrieval (Top k @ 10: 70% precision, 53% accuracy).

These results confirm that both RCS and the agentic architecture are indispensable components, while evidence integration depth requires careful calibration for multiple-choice based scientific research performance.

3.1.4 Impact of Language Model Choice Across System Components

The systematic evaluation of different language models across PaperQA2's components reveals performance variations that highlight the importance of model selection for scientific literature tasks. The component-wise analysis shows that individual PaperQA2 tools exhibit varying sensitivity to model upgrades, with the most dramatic improvements occurring when all components are integrated with stronger models.

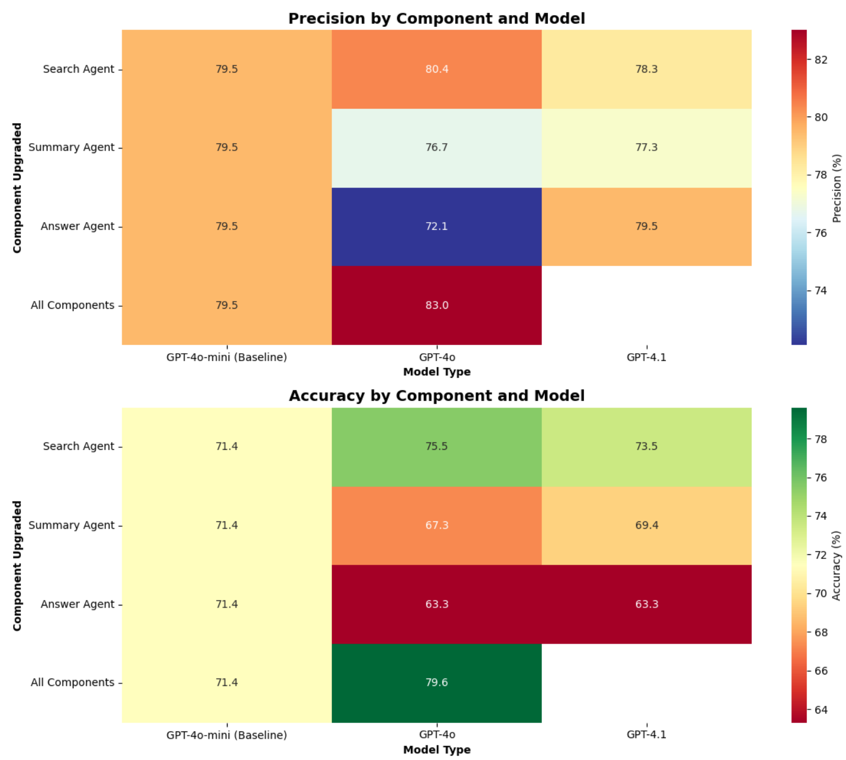

Figure 11: PaperQA2 Performance by Component and Model: Heatmaps showing (top) precision and (bottom) accuracy percentages for different PaperQA2 system components across three model types.

The Search Agent demonstrates consistency across model types, maintaining precision around 78-80% regardless of whether GPT-4o-mini, GPT-4o, or GPT-4.1 is used. That means, the paper search functionality is robust and less dependent on advanced reasoning capabilities. However, upgrading from the baseline to GPT-4o in the search component alone yields modest improvements (+0.9 precision, +4.1 accuracy), indicating that even this stable component benefits from enhanced model capabilities.

The Summary Agent shows moderate sensitivity to model choice, with precision ranging from 75-80% across different models, while the Answer Agent shows the highest variability

(72-80% in precision), suggesting that final response generation requires the most sophisticated reasoning capabilities. Furthermore, upgrading individual components in isolation can sometimes lead to performance degradation, as seen with the Summary → GPT-4o configuration (-2.8 precision, -4.1 accuracy) and Answer → GPT-4o configuration (-7.4 precision, -8.1 accuracy), suggesting that component interactions are complex and require balanced optimization.

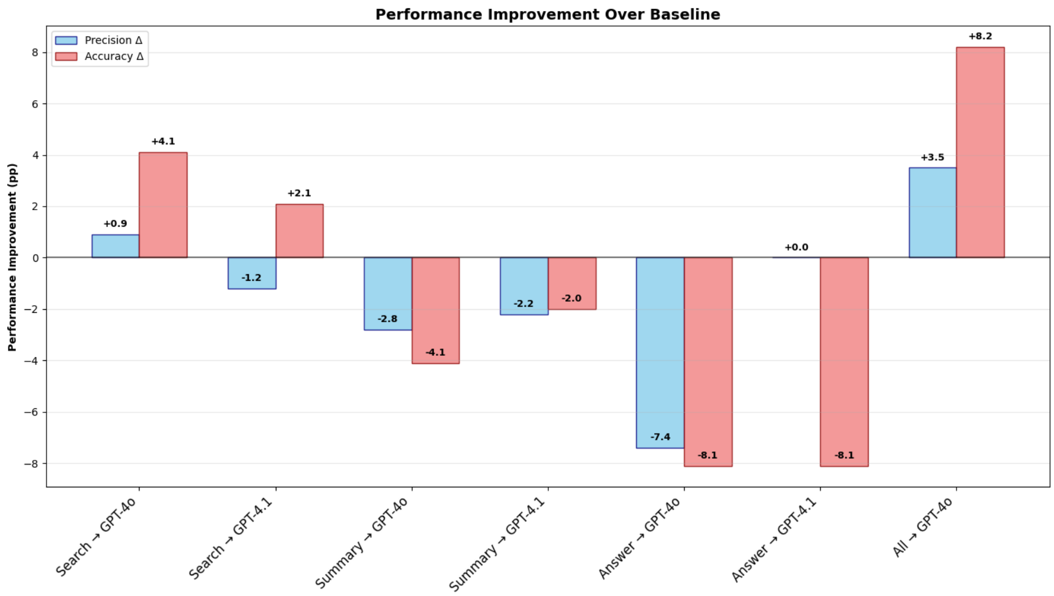

Figure 12: Performance Improvement Over Baseline: Comparison of precision and accuracy improvements (in percentage points) for different PaperQA2 model configurations relative to GPT-4o-mini baseline.

The most striking finding is from the All Components configuration, where integrating GPT-4o across all tools produces substantial performance gains (+3.5 precision, +8.2 accuracy), achieving 83% precision and 79.6% accuracy. That means, PaperQA2's agentic architecture benefits most from coherent model upgrades across all components rather than piecemeal improvements, with the integrated system substantially outperforming individual component upgrades and highlighting the importance of holistic system optimization for complex scientific literature tasks.

3.1.5 Chunking Strategy and Document Segmentation

Effective document chunking is critical for RAG systems, as it directly impacts the quality of retrieved context and subsequent answer generation. In RAG systems, large documents must be segmented into smaller, manageable pieces called chunks for efficient retrieval and processing. Two key parameters define this segmentation: chunk size (the maximum number of characters per chunk) and overlap (the number of characters shared between consecutive chunks to preserve context continuity).

We conducted a systematic evaluation of these chunking parameters to optimize PaperQA2's performance on the LitQA2 benchmark. We evaluated chunk size variations from 3,000 to 6,000 characters with fixed overlap of 250 characters, and overlap variations from 50 to 500 characters with fixed chunk size of 5,000 characters. All experiments used the text-embedding-3-small model with evidence_k=30 and max_sources=15, with performance measured using precision and accuracy.

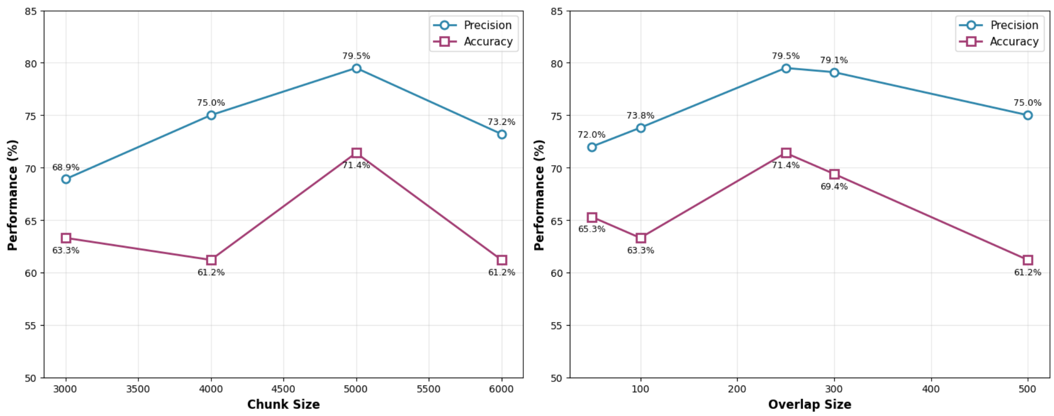

Figure 13: Chunking Strategy Optimization: Impact of Chunk Size and Overlap on LitQA2 Performance

Our chunk size analysis showed that the amount of text per chunk significantly affects both precision and accuracy. Small chunks of 3,000 characters achieved only 68.9% precision and 63.3% accuracy, likely due to insufficient contextual information for complex scientific questions that span multiple sentences or paragraphs. Performance improved substantially with larger chunks, peaking at 5,000 characters with 79.5% precision and 71.4% accuracy. However, further increases to 6,000 characters resulted in performance degradation to 73.2% precision and 61.2% accuracy, suggesting that overly large chunks introduce irrelevant information that dilutes the signal and complicates retrieval.

The overlap analysis revealed that the amount of shared text between consecutive chunks critically affects performance. Very small overlaps of 50-100 characters led to suboptimal performance (72.0-73.8% precision), likely because important information spanning chunk boundaries was fragmented, breaking contextual relationships essential for scientific comprehension. The optimal configuration was achieved with 250 characters overlap, maintaining peak performance of 79.5% precision and 71.4% accuracy. Larger overlaps of 300-500 characters showed diminishing returns, with 500 characters overlap dropping to 75.0% precision and 61.2% accuracy, possibly due to increased redundancy that introduces noise during retrieval and increases computational overhead.

Based on these results, we confirmed the optimal chunking strategy as 5,000 characters chunk size with 250 characters overlap (default values for PaperQA2). This configuration achieves the best balance between maintaining sufficient context within each retrievable unit while preserving information continuity across chunk boundaries. The 250-character overlap represents approximately 5% of the chunk size, providing adequate context preservation without excessive redundancy.

3.1.6 Summary and Implications for Domain-Specific Evaluation

Our comprehensive reproduction of PaperQA2 on the LitQA2 benchmark successfully validates the original findings, demonstrating that RAG agents can achieve human-level performance on scientific literature tasks. We show that: (1) the critical importance of both RCS and agentic architecture for optimal performance, (2) synergistic effects when upgrading all system components coherently, and (3) the need for careful parameter tuning in chunking strategies and evidence integration depth.

However, the LitQA2 evaluation, while rigorous, is constrained to biology domain multiple-choice questions where answers can be definitively verified against a single correct option.

This format, though valuable for systematic comparison, does not fully capture the complexity of real scientific inquiry workflows that researchers encounter in practice. Moreover, our ablation studies revealed that PaperQA2's RCS component, which proved beneficial for multiple-choice biology questions, may introduce information dilution effects that could differentially impact performance across scientific domains.

3.2 Evaluation Results on CosmoPaperQA2

Building on our LitQA2 reproduction findings, we now evaluate how different RAG architectures perform on the specialized demands of astronomical literature synthesis. Our evaluation of eight distinct RAG configurations (one without RAG as baseline) on 105 expert-curated cosmology questions reveals significant performance variations that show the challenges of scientific knowledge synthesis in astronomy compared to biology literature tasks.

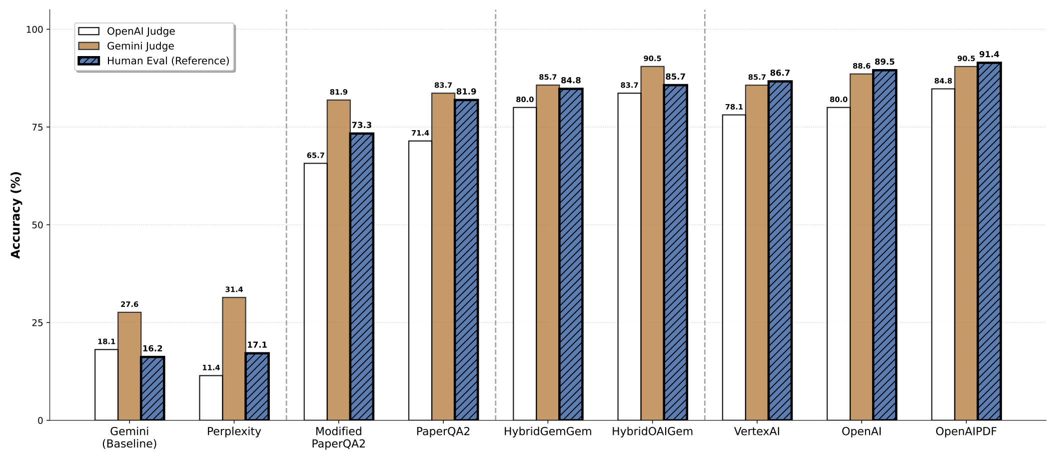

Figure 14: Performance comparison of SciRag Agents across three evaluation methods showing four distinct performance tiers. Vertical dashed lines separate the four performance groups. The first two entries (Gemini Baseline and Perplexity) do not perform RAG but simply rely on pre-trained LLM knowledge and, for Perplexity, built-in retrieval tools.

3.2.1 Human Evaluated Results

From the expert-evaluated results, we observe that the top-performing ones (OpenAIPDF, OpenAI, VertexAI) are all commercial RAGs, achieving 86.7-91.4% accuracy. Both hybrid implementations (HybridOAIGem: 85.7% , HybridGemGem: 84.8%) achieve performance competitive with commercial RAGs. PaperQA2 (81.90%) demonstrates solid performance but lags by 4.8-9.5 % compared to top performers. The poor performance of Perplexity Assistant (17.1%) and Gemini Assistant (16.2%) shows that unfiltered web search and non-RAG integration are insufficient for expert-level scientific inquiry, reinforcing the essential role of RAG Agents in scientific knowledge synthesis for autonomous scientific discovery.

While PaperQA2 demonstrates solid performance, its relative underperformance on the astronomy benchmark compared to its strong results on LitQA2 can be largely attributed to the diluting effect of its RCS component. The RCS step removes or compresses critical astronomical details, mathematical relationships, and technical specifics that are essential for accurate cosmology question answering. Furthermore, the performance variation between standard and Modified PaperQA2 configurations highlights the system's sensitivity to prompt design, where domain-specific optimization for astronomical terminology and response conciseness can significantly alter output quality and alignment with expert evaluation criteria.

Technical Analysis of Top Performers: OpenAI Assistants (89.5-91.4%) use OpenAI's file search tool, which combines automatic query rewriting, parallel searches, keyword and semantic search, and result reranking. This multi-faceted approach outperforms simple semantic-only retrieval used in hybrid systems (84.8-85.7%). Future work should evaluate domain-specific retrieval enhancements such as hybrid sparse-dense methods, contextual chunk expansion, query decomposition strategies, and multi-hop reasoning approaches to further optimize RAG performance for scientific applications.

3.2.2 AI Evaluated Results

Performance Tier Validation: Both OpenAI and Gemini judges successfully identify the four distinct tiers observed in human evaluation. The performance gaps are preserved: baseline systems achieve 11.4-18.1% (OpenAI judge) and 16.2-31.4% (Gemini judge), while top-performing agents reach 80.0-84.8% (OpenAI judge) and 88.6-91.4% (Gemini judge).

Judge-Specific Patterns: Our analysis shows systematic but predictable biases in AI evaluation that researchers must account for when using automated assessment. The OpenAI judge demonstrates consistent conservative scoring, rating systems 2-8% lower than human experts across all performance tiers. This conservative bias suggests the OpenAI judge applies stricter accuracy standards, potentially providing a more rigorous lower-bound estimate of system capabilities.

Conversely, the Gemini judge shows systematic overrating, consistently scoring systems 5-15 percentage points higher than human evaluation. Notable examples include Gemini Baseline (27.6% vs Human: 16.2%) and Modified PaperQA2 (81.9% vs Human: 73.3%). This optimistic bias may reflect the judge's tendency to credit partial correctness or tangentially related information, potentially overestimating real-world performance.

For researchers seeking robust performance estimates, the OpenAI judge's conservative scoring provides a safer lower bound for system capabilities, while Gemini's optimistic scoring may overestimate real-world performance. Despite these systematic biases, the consistent ranking order across all three evaluation methods (Pearson $r > 0.99$) validates the robustness of our assessment framework

3.2.3 Cost Performance Analysis

Our comprehensive cost-performance analysis reveals significant variations in efficiency across RAG implementations, with implications for both research applications and production deployment.

VertexAI emerges as the optimal choice for cost-conscious applications, demonstrating exceptional cost efficiency (86.7% accuracy) while maintaining strong performance. This efficiency comes from Gemini 2.5 Flash's competitive pricing structure: \$0.00015 per 1K input tokens and \$0.0006 per 1K output tokens for standard responses, representing a 13.3× cost advantage over GPT-4.1 for input tokens and a 13.3× advantage for output tokens.

OpenAI systems achieve the highest absolute performance (89.5-91.4% accuracy) but at substantial cost. GPT-4.1's pricing (\$0.002 per 1K input tokens, \$0.008 per 1K output tokens) reflects its advanced capabilities but may limit scalability for high-volume applications. The 2-5% performance advantage over VertexAI must be weighed against the significant cost for specific use cases.

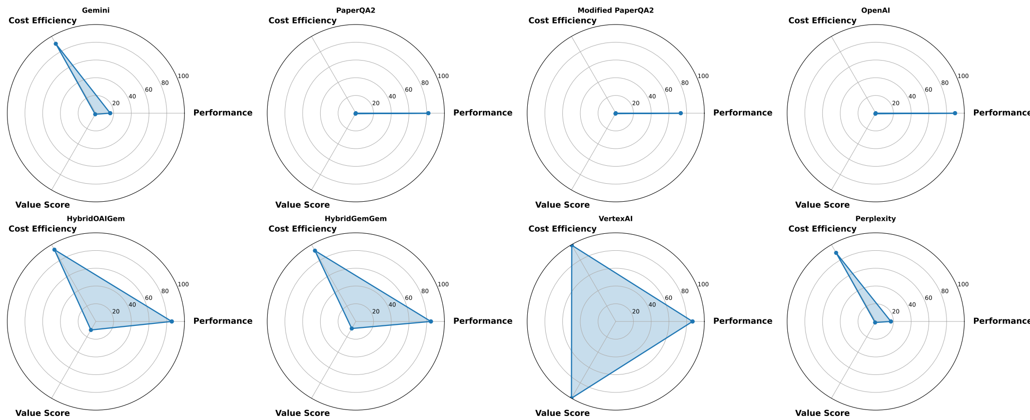

Figure 15: Multi-dimensional performance analysis of SciRag Agents across three key metrics: Performance (accuracy score), Cost Efficiency (inverse of operational cost), and Value Score (performance per unit cost). Each radar chart represents one agent, with larger areas indicating better overall value. Cost estimates are approximated using identical queries across different SciRag Agents for comparison.

The hybrid approaches (HybridOAIGem, HybridGemGem) offer compelling balanced trade-offs, achieving 84.8-85.7% accuracy while maintaining moderate operational costs. These systems demonstrate that strategic combination of high-quality embeddings with cost-effective generation models can approach commercial performance at reduced expense, while avoiding the information loss associated with intermediate summarization steps, demonstrated by PaperQA2.

Both Gemini and Perplexity show limited performance across all evaluated dimensions, with particularly poor cost-efficiency ratios.

4 Discussion

4.1 Reproduction Validation and Implementation Considerations

Our reproduction successfully validates the core claims from the original PaperQA2 work, demonstrating that RAG agents can achieve superhuman performance on scientific literature tasks. However, several implementation differences provide important insights into system robustness and evaluation methodology. The most significant limitation was the unavailability of the Citation Traversal tool, which constrained our reproduction scope to exclude DOI recall analysis. Additionally, our single-run experiments, while computationally more efficient than the original three-run approach, prevented direct statistical comparison with reported confidence intervals.

Despite these methodological differences, our results align closely with the original findings. This consistency suggests that PaperQA2's core architecture is robust to reasonable implementation variations, though the performance variations between our training and test datasets highlight the importance of data quality and temporal consistency in benchmark design.

4.2 Domain-Specific RAG Optimization

The most significant discovery from our cross-domain evaluation is that different scientific domains require fundamentally different RAG configurations. This finding challenges the prevalent assumption of universal RAG architectures.

Our observations align with recent theoretical frameworks for understanding query complexity in data-augmented LLM applications. Zhao et al. (2024)[33] propose a comprehensive taxonomy that categorizes queries into four levels based on the type of external data required and reasoning complexity: explicit facts (Level-1), implicit facts (Level-2), interpretable rationales (Level-3), and hidden rationales (Level-4).

The RCS component exemplifies this domain sensitivity. While RCS proved essential for biology literature tasks in LitQA2—improving precision from 68% to 80% on the test set—our astronomy evaluation suggests it may actually be detrimental for cosmological literature synthesis. According to Zhao et al.'s framework, biology questions in LitQA2 primarily represent Level-1 and Level-2 queries (explicit and implicit fact retrieval), where RCS's information compression and filtering successfully eliminates irrelevant details while preserving factual content. However, astronomical research queries often require Level-3 interpretable rationales involving mathematical relationships, observational constraints, and theoretical frameworks that the RCS step may inadvertently filter out or compress, removing critical domain-specific details essential for accurate reasoning.

This domain sensitivity extends beyond individual components to entire system architectures. Commercial systems (OpenAI Assistant: 91.4%, VertexAI: 86.7%) significantly outperformed PaperQA2 (81.9%) on astronomical tasks. In contrast, PaperQA2's semantic-only approach, optimized for biology literature, may be insufficient for the mathematical and observational complexity characteristic of astronomical research.

These findings fundamentally challenge one-size-fits-all RAG approaches and support Zhao et al.'s assertion that there is no one-size-fits-all solution for data-augmented LLM applications. Our empirical evidence demonstrates that domain-specific optimization is not merely beneficial but essential for achieving optimal performance in scientific applications, with the choice of RAG architecture critically dependent on both the domain's characteristic query types and the underlying reasoning complexity required for accurate knowledge synthesis.

4.3 Limitations and Future Directions

Several limitations constrain the generalizability of our findings. First, our evaluation spans only two scientific domains (biology and astronomy), limiting conclusions about RAG optimization across the broader scientific landscape. Future work should systematically evaluate RAG architectures across diverse scientific fields—chemistry, physics, materials science—to establish domain-specific optimization principles.

Second, our reproduction was constrained by missing components (Citation Traversal tool) and computational limitations (single-run experiments). While our results align with original findings, complete reproduction would strengthen confidence in reported performance claims and enable more precise statistical comparison.

Third, current evaluation frameworks remain limited. The LitQA2 benchmark, while rigorous, constrains evaluation to multiple-choice biology questions. CosmoPaperQA2 addresses this with open-ended astronomical questions, but broader benchmarks spanning diverse scientific domains and query types are needed for comprehensive RAG assessment.

Technical limitations include the need for more sophisticated evaluation metrics beyond accuracy and precision. Future frameworks should assess reasoning quality, source integration, uncertainty quantification, and response reliability—critical capabilities for autonomous scientific discovery applications.

4.4 Implications for Scientific AI Integration

Our findings have significant implications for integrating AI systems into scientific research workflows. The demonstrated domain sensitivity suggests that scientific institutions should invest in domain-specific RAG optimization rather than adopting generic solutions. This has practical implications for research computing infrastructure, requiring domain expertise during system design and deployment.

For autonomous scientific discovery applications, our framework enables systematic selection of optimal RAG configurations based on specific research requirements. The calibrated AI evaluators provide scalable assessment tools capable of handling thousands of scientific queries, essential for large-scale autonomous research systems.

Most importantly, our work demonstrates that RAG agents have matured to achieve expert-level performance on domain-specific scientific tasks. This capability, combined with systematic optimization frameworks, positions AI systems to serve as genuine research partners rather than mere information retrieval tools. However, successful integration requires careful attention to domain-specific requirements, evaluation methodology, and cost-performance trade-offs—insights that will guide the next generation of scientific AI systems.

5 Conclusion

This work successfully validates and extends the foundational claims of PaperQA2, that is, RAG agents can indeed achieve superhuman performance on scientific literature tasks. However, our cross-domain evaluation reveals an insight that challenges prevailing assumptions about universal RAG architectures: different scientific domains require fundamentally different optimization strategies.

Our reproduction of PaperQA2 on LitQA2 confirms the original findings, with our implementation achieving 71% accuracy on the test set while validating the critical importance of both the RCS component and agentic architecture for biology literature tasks. The systematic ablation studies further demonstrate that coherent model upgrades across all system components yield superior performance compared to piecemeal improvements.

The extension to astronomical literature through our SciRag framework and CosmoPaperQA2 benchmark reveals interesting domain-specific performance variations that have profound implications for scientific AI deployment. While PaperQA2 excels in biology literature synthesis, commercial RAG systems significantly outperform it in astronomy (91.4% vs 81.9%), suggesting that domain-specific architectural choices are not merely optimization opportunities but fundamental requirements for achieving expert-level performance.

The practical implications extend beyond academic interest to real-world deployment considerations. Our cost-performance analysis provides actionable guidance for research institutions, identifying VertexAI as optimal for cost-conscious applications while highlighting that OpenAI systems achieve the highest absolute performance at premium costs. The calibrated AI evaluators we developed enable scalable assessment of thousands of scientific queries, essential for autonomous research applications.

Looking forward, these findings position the field to move beyond generic RAG solutions toward domain-aware architectures that adapt their processing strategies based on the underlying scientific domain and query complexity. As RAG agents mature to serve as genuine research partners rather than information retrieval tools, the framework and insights presented here provide essential guidance for the next generation of scientific AI systems that can truly support autonomous scientific discovery across diverse research domains.

A RAG Prompts

Our modified PaperQA2 prompt priorities conciseness and domain specificity for efficient human evaluation.

Modified PaperQA2 Prompt

Provide a concise answer in 1-2 sentences maximum.
Context (with relevance scores):{context}
Question: {question}
Write a concise answer based on the context, focusing on astronomical facts and concepts. If the context provides insufficient information, reply {CANNOT_ANSWER_PHRASE}.
Write in the style of a scientific astronomy reference, with precise and factual statements. The context comes from a variety of sources and is only a summary, so there may be inaccuracies or ambiguities.
{prior_answer_prompt} Answer (maximum one sentence):

In contrast, the original prompt emphasizes comprehensive information synthesis, mandatory citation and Wikipedia-style formatting.

PaperQA2 Prompt

Answer the question below with the context.

Context (with relevance scores):{context}
Question: {question}
Write an answer based on the context.
If the context provides insufficient information reply {CANNOT_ANSWER_PHRASE}
For each part of your answer, indicate which sources most support it via citation keys at the end of sentences, like {example_citation}.
Only cite from the context above and only use the citation keys from the context.
{CITATION_KEY_CONSTRAINTS}
Do not concatenate citation keys, just use them as is.

Write in the style of a Wikipedia article, with concise sentences and coherent paragraphs. The context comes from a variety of sources and is only a summary, so there may inaccuracies or ambiguities. If quotes are present and relevant, use them in the answer. This answer will go directly onto Wikipedia, so do not add any extraneous information.
{prior_answer_prompt}
Answer ({answer_length}):

The Hybrid SciRag assistant adopt a structured approach, requiring a JSON format return for consistent response parsing.

Hybrid Assistants Prompt

You are a helpful assistant. Answer based on the provided context. You must respond in valid JSON format with the following structure:

{ "answer": "your detailed answer here", "sources": ["source1", "source2", "source3"]}

The sources must be from the **Context** material provided. Include source names, page numbers, equation numbers, table numbers, section numbers when available. Ensure your response is valid JSON only.

The Perplexity assistant uses web search to specific papers while utilizing its real-time retrieval capabilities.

Perplexity Assistants Prompt

You are a scientific literature search agent specializing in cosmology.
We perform retrieval on the following set of papers: {paper_list}
Your task is to answer questions using ONLY information from these specific papers. Do not use any other sources or general knowledge beyond what these papers contain.

Instructions:
1. Search for information relevant to the question within the specified papers
2. Provide a CONCISE answer in EXACTLY 1-3 sentences. Do not exceed 3 sentences under any circumstances.
3. Add numerical references [1], [2], [3], etc. corresponding to the paper numbers listed above
4. If the papers don't contain sufficient information, state this clearly in 1-2 sentences maximum
5. Focus ONLY on the most important quantitative results or key findings
6. Be precise, direct, and avoid any unnecessary elaboration or context

CRITICAL: Your answer section must contain no more than 3 sentences total. Count your sentences carefully.
You must search your knowledge base calling your tool. The sources must be from the retrieval only.
Your response must be in JSON format with exactly these fields:
- "answer": Your 1-3 sentence response with citations
- "sources": Array of citation numbers used (e.g., ["1", "2"])

Gemini Assistant's approach to leveraging pre-trained knowledge of specific cosmological papers without requiring external retrieval mechanisms.

Gemini Assistant Prompt

You are a scientific literature agent specializing in cosmology.
You have access to the following key cosmology papers in your knowledge base: {paper_list}
Your task is to answer cosmology questions using your knowledge of these papers and general cosmology knowledge. Instructions: 1. Answer the question based on your knowledge of cosmology and the listed papers
2. Provide a CONCISE answer in EXACTLY 1-2 sentences maximum
3. Add numerical references [1], [2], [3], etc. when citing the specific papers listed above
4. Focus ONLY on the most important quantitative results or key findings
5. Be precise, direct, and avoid any unnecessary elaboration

Paper reference guide:
[1] - Planck 2018 cosmological parameters
[2] - CAMELS machine learning cosmology simulations
[3] - Single galaxy cosmology analysis
[4] - Local Hubble constant measurement (Riess et al.)
[5] - Atacama Cosmology Telescope DR6 results

CRITICAL: Your answer must be no more than 2 sentences total. Count your sentences carefully.
Your response must be in JSON format with exactly these fields:
- "answer": Your 1-2 sentence response with citations
- "sources": Array of paper citations [1]-[5] that are relevant to your answer

OpenAI/VertexAI assistants use a tool-based retrieval approach with markdown formatting, emphasising precise source and knowledge integration.

OpenAI/Vertex Assistants Prompt

You are a retrieval agent. You must add precise source from where you got the answer. Your answer should be in markdown format with the following structure:
Answer:{answer}
Sources:{sources}
You must search your knowledge base calling your tool. The sources must be from the retrieval only. You must report the source names in the sources field, if possible, the page number, equation number, table number, section number, etc.

B Chain of Thought

AI judges are given the following prompt:

Judge Prompt

You are an expert scientific evaluator assessing the quality of scientific responses against reference answers.

Your task is to evaluate responses using one critical criterion:
ACCURACY (0-100):
CRITICAL: Use ONLY these two scores for accuracy:
- 100: The answer contains the core correct factual content, concepts, and conclusions from the ideal answer
- 0: The answer is fundamentally wrong or contradicts the ideal answer
This is a BINARY evaluation - either the answer is essentially correct (100) or fundamentally incorrect (0).
No partial credit or intermediate scores allowed.

EVALUATION GUIDELINES:
- Focus ONLY on whether the main scientific concepts and conclusions are correct
- Check that the core factual claims from the ideal answer are present in the generated answer
- Verify the overall conceptual direction and main conclusions align
- Additional correct information beyond the ideal answer is acceptable
- Only award 0 if the answer contradicts the ideal answer or gets the main concepts wrong
- Award 100 if the answer captures the essential correct scientific understanding

Provide your evaluation with the numerical score and detailed rationale explaining why you chose 100 or 0."""

Please evaluate this system's response against the ideal answer:
QUESTION: {question}
GENERATED ANSWER:
{generated_answer}
IDEAL ANSWER:
{ideal_answer}

Evaluate based on:
Accuracy (0-100): How factually correct is the answer compared to the ideal?
Use the evaluate_response function to provide your structured evaluation with detailed rationale.

C AG2 Prompt

The following is the AG2 prompt to extract multiple choice answer from PaperQA2 Agents.

AG2 Prompt

Your task is to analyze PaperQA2 responses to multiple-choice questions and determine which answer option is being selected by the model. Be precise and objective.

You will receive:

1. A question with multiple-choice options

2. The model's response

For each response, you must:

1. Analyze the content carefully

2. Determine which option the response is selecting

3. Return ONLY the single letter of the selected option (A, B, C, D, etc.)

4. If the model is indicating it doesn't have enough information, select the "Insufficient information" option

IMPORTANT: Your response must contain EXACTLY ONE LETTER, nothing else.

IMPORTANT RULES FOR NUMERICAL QUESTIONS:

- When the question involves numerical values, pay attention to significant figures

- If PaperQA2's response contains a numeric value with higher precision than the options (e.g., 45.67% vs 46%), round to the same number of significant figures as in the options

- Match to the closest option after appropriate rounding

- If two options are equally close after rounding, choose the one that appears in the response

- If the response is significantly different from all options, select the "Insufficient information" option Do not include explanations, punctuation, or any other text.

References

[1] 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.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds., Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. [Online]. Available: https://aclanthology.org/D15-1075/.

[3] J. Genesis and F. Keane, "Integrating knowledge retrieval with generation: A comprehensive survey of rag models in nlp," *Preprints*, Apr. 2025. DOI: 10.20944/preprints202504.0351.v1. [Online]. Available: https://doi.org/10.20944/preprints202504.0351.v1.

[4] S. Zeng, J. Zhang, P. He, *et al.*, *The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)*, 2024. arXiv: 2402.16893 [cs.CR]. [Online]. Available: https://arxiv.org/abs/2402.16893.

[5] J. Li, Y. Yuan, and Z. Zhang, *Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases*, 2024. arXiv: 2403.10446 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.10446.

[6] X. Wang, Z. Wang, X. Gao, *et al.*, *Searching for best practices in retrieval-augmented generation*, 2024. arXiv: 2407.01219 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2407.01219.

[7] V. Karpukhin, B. Oguz, S. Min, *et al.*, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550. [Online]. Available: https://aclanthology.org/2020.emnlp-main.550/.

[8] K. Sawarkar, A. Mangal, and S. R. Solanki, "Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2024, pp. 155–161. DOI: 10.1109/MIPR62202.2024.00031.

[9] S. Chen, B. J. Gutiérrez, and Y. Su, *Attention in large language models yields efficient zero-shot re-rankers*, 2025. arXiv: 2410.02642 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2410.02642.

[10] Anthropic, *Introducing contextual retrieval*, Accessed June 2, 2025, 2024. [Online]. Available: https://www.anthropic.com/news/contextual-retrieval.

[11] P. Lewis, E. Perez, A. Piktus, *et al.*, *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021. arXiv: 2005.11401 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.11401.

[12] A. Asai, J. He, R. Shao, *et al.*, *Openscholar: Synthesizing scientific literature with retrieval-augmented lms*, 2024. arXiv: 2411.14199 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2411.14199.

[13] I. Ciucă, Y.-S. Ting, S. Kruk, and K. Iyer, *Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy*, 2023. arXiv: 2306.11648 [astro-ph.IM]. [Online]. Available: https://arxiv.org/abs/2306.11648.

[14] W. Shao, P. Ji, D. Fan, *et al.*, *Astronomical knowledge entity extraction in astrophysics journal articles via large language models*, 2024. arXiv: 2310.17892 [astro-ph.IM]. [Online]. Available: https://arxiv.org/abs/2310.17892.

[15] J. F. Wu *et al.*, *Designing an evaluation framework for large language models in astronomy research*, 2024. arXiv: 2405.20389 [astro-ph.IM]. [Online]. Available: https://arxiv.org/abs/2405.20389.

[16] Y.-S. Ting *et al.*, *Astromlab 1: Who wins astronomy jeopardy!?* 2024. arXiv: 2407.11194 [astro-ph.IM]. [Online]. Available: https://arxiv.org/abs/2407.11194.

[17] J. Li, F. Zhao, P. Chen, *et al.*, "An astronomical question answering dataset for evaluating large language models," *Scientific Data*, vol. 12, p. 447, 2025. DOI: 10.1038/s41597-025-04613-9. [Online]. Available: https://doi.org/10.1038/s41597-025-04613-9.

[18] X. Zhong, B. Jin, S. Ouyang, *et al.*, *Benchmarking retrieval-augmented generation for chemistry*, 2025. arXiv: 2505.07671 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2505.07671.

[19] M. Yasunaga, J. Kasai, R. Zhang, *et al.*, *Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks*, 2019. arXiv: 1909.01716 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1909.01716.

[20] A. Dimitrov and B. Bolliet, *Cosmopaperqa*, https://huggingface.co/datasets/ASTROANTS/CosmoPaperQA, CosmoPaperQA, v1.0, 2025.

[21] N. Aghanim *et al.*, "Planck2018 results: Vi. cosmological parameters," *Astronomy and Astrophysics*, vol. 641, A6, Sep. 2020, ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833910. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201833910.

[22] F. Villaescusa-Navarro *et al.*, "The camels project: Cosmology and astrophysics with machine-learning simulations," *The Astrophysical Journal*, vol. 915, no. 1, p. 71, Jul. 2021, ISSN: 1538-4357. DOI: 10.3847/1538-4357/abf7ba. [Online]. Available: http://dx.doi.org/10.3847/1538-4357/abf7ba.

[23] F. Villaescusa-Navarro *et al.*, "Cosmology with one galaxy?" *The Astrophysical Journal*, vol. 929, no. 2, p. 132, Apr. 2022, ISSN: 1538-4357. DOI: 10.3847/1538-4357/ac5d3f. [Online]. Available: http://dx.doi.org/10.3847/1538-4357/ac5d3f.

[24] A. G. Riess, L. M. Macri, S. L. Hoffmann, *et al.*, "A 2.4% determination of the local value of the Hubble constant," *The Astrophysical Journal*, vol. 826, no. 1, p. 56, Jul. 2016, ISSN: 1538-4357. DOI: 10.3847/0004-637x/826/1/56. [Online]. Available: http://dx.doi.org/10.3847/0004-637X/826/1/56.

[25] E. Calabrese *et al.*, *The atacama cosmology telescope: Dr6 constraints on extended cosmological models*, 2025. arXiv: 2503.14454 [astro-ph.CO]. [Online]. Available: https://arxiv.org/abs/2503.14454.

[26] OpenAI, *New embedding models and api updates*, 2023. [Online]. Available: https://openai.com/blog/new-embedding-models-and-api-updates.

[27] Google DeepMind, *Gemini pro*, https://deepmind.google/models/gemini/pro/, Accessed: 2025-01-XX, 2025.

[28] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2014–2037. DOI: 10.18653/v1/2023.eacl-main.148. [Online]. Available: https://aclanthology.org/2023.eacl-main.148/.

[29] UK AI Security Institute, *Inspect: A framework for large language model evaluations*, https://github.com/UKGovernmentBEIS/inspect_ai, Software framework for LLM evaluations with built-in components for prompt engineering, tool usage, multi-turn dialog, and model graded evaluations, 2024. [Online]. Available: https://inspect.aisi.org.uk/.

[30] C. Wang, Q. Wu, and the AG2 Community, *Ag2: Open-source agentos for ai agents*, version latest, Available at https://docs.ag2.ai/, 2024. [Online]. Available: https://github.com/ag2ai/ag2.

[31] J. Gu, X. Jiang, Z. Shi, *et al.*, *A survey on llm-as-a-judge*, 2025. arXiv: 2411.15594 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2411.15594.

[32] S. Dai, Y. Zhou, L. Pang, *et al.*, "Neural retrievers are biased towards llm-generated content," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '24, Barcelona, Spain: Association for Computing Machinery, 2024, pp. 526–537, ISBN: 9798400704901. DOI: 10.1145/3637528.3671882. [Online]. Available: https://doi.org/10.1145/3637528.3671882.

[33] S. Zhao, Y. Yang, Z. Wang, Z. He, L. K. Qiu, and L. Qiu, *Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely*, 2024. arXiv: 2409.14924 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2409.14924.