

PaperScope: A Multi-Modal Multi-Document Benchmark for Agentic Deep Research Across Massive Scientific Papers

Lei Xiong^{1,2}, Huaying Yuan¹, Zheng Liu^{2*}, Zhao Cao¹, Zhicheng Dou^{1*}

¹Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

²Beijing Academy of Artificial Intelligence, Beijing, China

xiongxiongleilei@ruc.edu.cn

Abstract

Leveraging Multi-modal Large Language Models (MLLMs) to accelerate frontier scientific research is promising, yet how to rigorously evaluate such systems remains unclear. Existing benchmarks mainly focus on single-document understanding, whereas real scientific workflows require integrating evidence from multiple papers, including their text, tables, and figures. As a result, multi-modal, multi-document scientific reasoning remains underexplored and lacks systematic evaluation. To address this gap, we introduce PaperScope, a multi-modal multi-document benchmark designed for agentic deep research. PaperScope presents three advantages: (1) **Structured scientific grounding.** It is built on a knowledge graph of over 2,000 AI papers spanning three years, providing a structured foundation for research-oriented queries. (2) **Semantically dense evidence construction.** It integrates semantically related key information nodes and employs an optimized random-walk article selector to sample thematically coherent paper sets, thereby ensuring adequate semantic density and task complexity. (3) **Multi-task evaluation of scientific reasoning.** It contains over 2,000 QA pairs across reasoning, retrieval, summarization, and problem solving, enabling evaluation of multi-step scientific reasoning. Experimental results show that even advanced systems such as OpenAI Deep Research and Tongyi Deep Research achieve limited scores on PaperScope, highlighting the difficulty of long-context retrieval and deep multi-source reasoning. PaperScope thus provides a rigorous benchmark alongside a scalable pipeline for constructing large-scale multi-modal, multi-source deep research datasets. The code and dataset are available at: <https://github.com/CherYou/PaperScope>.

1 Introduction

Scientific papers are inherently multi-modal, and solving complex scientific problems demands re-

trieving, reasoning, and synthesizing information across multiple documents that span text, tables, figures, formulas, and algorithms. Agentic deep research systems (OpenAI, 2025a; Google, 2025a; Team et al., 2025c) have reshaped research workflows by autonomously planning retrieval strategies, invoking external tools, refining queries adaptively, and verifying context, thereby exhibiting stronger dynamic reasoning (Geng et al., 2025; Li et al., 2025a; Jin et al., 2025) than traditional RAG for research-style tasks.

However, current systems remain primarily grounded in internet text resources and often overlook large volumes of stored multi-modal documents such as papers, technical reports, financial reports, and lab manuals that encode dense, structured visual and symbolic information (Dong et al., 2025); failing to leverage these artifacts limits the applicability and knowledge coverage in realistic scientific pipelines. Existing benchmarks also fall short: document understanding datasets mainly evaluate single-page or multi-page single-document tasks (Masry et al., 2022; Mathew et al., 2021; Cui et al., 2025; Mathew et al., 2022; Li et al., 2024c; Ma et al., 2024; Li et al., 2024b,a; Tian et al., 2025; Li et al., 2025b; Wang et al., 2024), and agent benchmarks largely target general-purpose assistants and web browsing (Wei et al., 2025b; Mialon et al., 2023a; Phan et al., 2025). As a result, multi-modal, multi-document scientific reasoning, especially in large-file environments, remains underrepresented, and there is a pressing need for realistic, reproducible protocols that faithfully reflect research workflows.

To address these gaps, we introduce PaperScope, a multi-modal, multi-document agentic benchmark for scientific deep research. PaperScope introduces the following key features:

(1) **Assess agents' capabilities in multi-source information retrieval, synthesis, reasoning, and generation.** Each question is associated with a

*Corresponding author

large corpus, where answers are dispersed across heterogeneous sources and modalities. The benchmark corpus contains 202–1500 documents spanning multiple sub-fields of artificial intelligence, including reinforcement learning, generative models, and computer vision—thus covering a broad range of scientific research scenarios in AI.

(2) **Provide diverse task formats across scientific deep-research skills.** It includes Reasoning, Topic Induction, Summary, and Solution tasks, collectively capturing document retrieval, document understanding, multi-source information integration, and methodological formulation. The questions include both objective and subjective types; this diversity increases the evaluation difficulty and more faithfully reflects the complexity of real-world research applications.

(3) **Ensure high-quality, graph-grounded annotations.** Following an inverted-construction strategy, we build each question using document sets sampled through random walks on a knowledge graph constructed from top AI conferences, ensuring answer accuracy and near-uniqueness within the corpus-level knowledge graph. We further perform strict quality control on both documents and annotations to guarantee task specificity and correctness across all categories.

We conduct comprehensive experiments on PaperScope, evaluating two categories of models: standard ReAct-based agents and Deep Research agents for a total of 16 systems. Even leading open-source and closed-source agents fail to achieve strong performance. Our benchmark reveals that existing models still face substantial limitations in large-scale document deep-research scenarios: their accuracy remains insufficient, and their multi-modal understanding, retrieval, and multi-source information integration capabilities require significant improvement.

In summary, our contributions are threefold:

(1) We introduce PaperScope, a multi-modal, multi-document benchmark for scientific deep research, targeting large-scale retrieval, cross-document reasoning, and multi-source information synthesis under realistic research settings.

(2) PaperScope is built from a heterogeneous corpus via a knowledge-graph-guided construction pipeline, providing fine-grained, near-unique annotations and a reproducible methodology for creating multi-source, multi-modal, multi-document deep-research datasets.

(3) Experiments on 16 state-of-the-art ReAct-

based and Deep Research agents reveal substantial performance gaps in accuracy, multi-modal understanding, retrieval, and information integration, highlighting the difficulty of large-scale scientific deep research.

2 Related Work

Long Scientific Document Understanding Benchmarks. Existing benchmarks for long scientific document understanding primarily focus on single-document settings and remain limited in cross-page or multi-source reasoning. Charxiv (Wang et al., 2024) and ArxivQA (Li et al., 2024b) construct figure-centric QA with limited grounding in full-paper content, while MMSCI (Li et al., 2024c) is restricted to caption-based tasks. CURIE (Cui et al., 2025) extends to scientific problem solving across multiple domains but still operates within individual documents. MMCR (Tian et al., 2025) and M3SciQA (Li et al., 2024a) introduce cross-source or citation-based settings, yet do not support genuine multi-document synthesis. DeepSolution (Li et al., 2025b) evaluates solution generation from structured PDF content, and M4DocBench (Dong et al., 2025) targets multi-modal and multi-turn reasoning under limited human annotations, but neither emphasizes multi-document retrieval or complex problem solving. Overall, existing datasets remain largely confined to single papers or single sources, leaving cross-page, multi-modal, and multi-document scientific retrieval, reasoning, summarization, and problem solving insufficiently explored.

General Agentic Deep Research Benchmarks. Agentic deep research systems extend traditional RAG through autonomous retrieval planning, adaptive tool use, and contextual verification. However, existing agent benchmarks—such as HLE, BrowseComp, and GAIA (Phan et al., 2025; Wei et al., 2025b; Mialon et al., 2023b)—primarily target web-based general-purpose information seeking and offer limited coverage of scientific documents rich in symbolic, visual, and algorithmic structures. Their tasks rarely require integrating multiple local files or synthesizing multi-modal evidence, leaving key aspects of real scientific workflows untested. PaperScope Bench addresses these gaps by constructing high-correlation scientific subsets and emphasizing cross-document retrieval, multi-modal grounding, and multi-resource reasoning, enabling a unified evaluation of retrieval,

| Benchmarks | Avg. Papers | Avg. Pages | Cross-Page | Multi-Modal | Multi-Doc | Capability | Open Form | Agent |
|--|-----------------|-----------------|------------|-------------|-----------|--------------------------------------|-----------|-------|
| LSDU Benchmark | | | | | | | | |
| DocVQA | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| ChartQA | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| InfoVQA | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| Charxiv | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| ArxivQA | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| MMSCI | - | 1 | ✗ | ✗ | ✗ | U | ✗ | ✗ |
| MMLongBench-Doc | 1 | 47.5 | ✓ | ✗ | ✗ | U, SRea | ✗ | ✗ |
| MMCR | 1 | 19 | ✓ | ✓ | ✗ | U, MRea | ✗ | ✗ |
| CURIE | 1 | - | ✓ | ✓ | ✗ | U | ✗ | ✗ |
| M3SciQA | 2 | - | ✓ | ✓ | ✓ | Retr, U | ✗ | ✗ |
| DeepSolution | 1 | - | ✓ | ✗ | ✗ | U | ✓ | ✗ |
| Agentic Deep Research Benchmark | | | | | | | | |
| HLE | - | - | - | ✗ | ✗ | Rea, Think, G | ✗ | ✓ |
| BrowseComp | - | - | - | ✗ | Web | Retr, MRea, Think | ✗ | ✓ |
| GAIA | - | - | - | ✗ | Web | Retr, MRea, Solu, Think, G | ✓ | ✓ |
| M4DocBench | 3.8 | 7 | ✓ | ✓ | ✓ | Retr, U, MRea, Think, G | ✓ | ✓ |
| PaperScope Bench | >=500 | >5000 | ✓ | ✓ | ✓ | Retr, U, MRea, Solu, Think, G | ✓ | ✓ |

Table 1: Comparison of existing benchmarks and our proposed **Agentic Deep Research Benchmark**. Symbols: ✓ indicates support, ✗ indicates lack of support. Abbreviations: U = Understanding, SRea = Single-resource reasoning, MRea = Multi-resource reasoning, Retr = Retrieval, Solu = Solution, Think = Thinking, G = Generation.

understanding, reasoning, and solution generation.

3 PaperScope Bench

3.1 Overview

We propose PaperScope, a multi-modal benchmark designed to evaluate Deep Research agents across a large-scale scientific corpus. As outlined in Table 2 and Figure 1, the benchmark comprises 2,400 questions spanning 11 sub-tasks. These are systematically organized into four meta-categories—*Topic Induction*, *Multi-Document Reasoning*, *Summary*, and *Solution*—which are structured to mirror the hierarchical real-world research workflow from retrieval to synthesis. Detailed task statistics are provided in Appendix A.

3.2 Data Collection

To evaluate retrieval, scientific understanding, synthesis, and problem solving across a large-scale scientific corpus, we construct a multi-modal, multi-document deep-research dataset built from scientific papers. Existing benchmarks rarely test complex multi-document retrieval and reasoning, motivating our new design. We collect 25,495 papers from ArXiv¹ and OpenReview², covering top AI conferences from 2023–2025 across more than 20 venues. All PDFs are rigorously filtered for completeness, readability, and quality, ensuring high data integrity and domain relevance. This curation allows PaperScope Bench to capture the scale, diversity, and complexity of real scientific research

¹arXiv: <https://arxiv.org>

²openreview: <https://openreview.net>

scenarios.

3.3 Task Creation

Topic Induction Tasks. This task evaluates a model’s ability to perform multi-modal semantic retrieval and topic recognition within a large scientific corpus. Given a textual or multi-modal query, the model must retrieve relevant papers under two settings: (1) *Implicit Topic Induction*, where the query provides latent semantic cues and the model must infer the underlying theme; (2) *Explicit Topic Induction*, where the topic is directly specified and the model must identify the most relevant studies.

Reasoning Tasks. This task measures cross-document reasoning and multi-modal evidence integration—capabilities central to scientific deep research. After retrieving related papers, the model must integrate textual and visual information across sources. Five sub-tasks are included: (1) *Figure–Table–Chart Comparison*, comparing core visual and quantitative findings across papers; (2) *Figure–Table–Chart Reasoning*, integrating diverse heterogeneous visual evidence for logical inference; (3) *Formula Reasoning*, interpreting mathematical expressions and key variable relationships; (4) *Algorithm Reasoning*, understanding pseudocode and core procedural logic; (5) *Full-Paper Reasoning*, synthesizing semantic and structural information across complete full texts.

Summary Tasks. This task assesses the ability to integrate knowledge across multiple papers and produce higher-level scientific abstractions. Given a thematic area and related documents, the

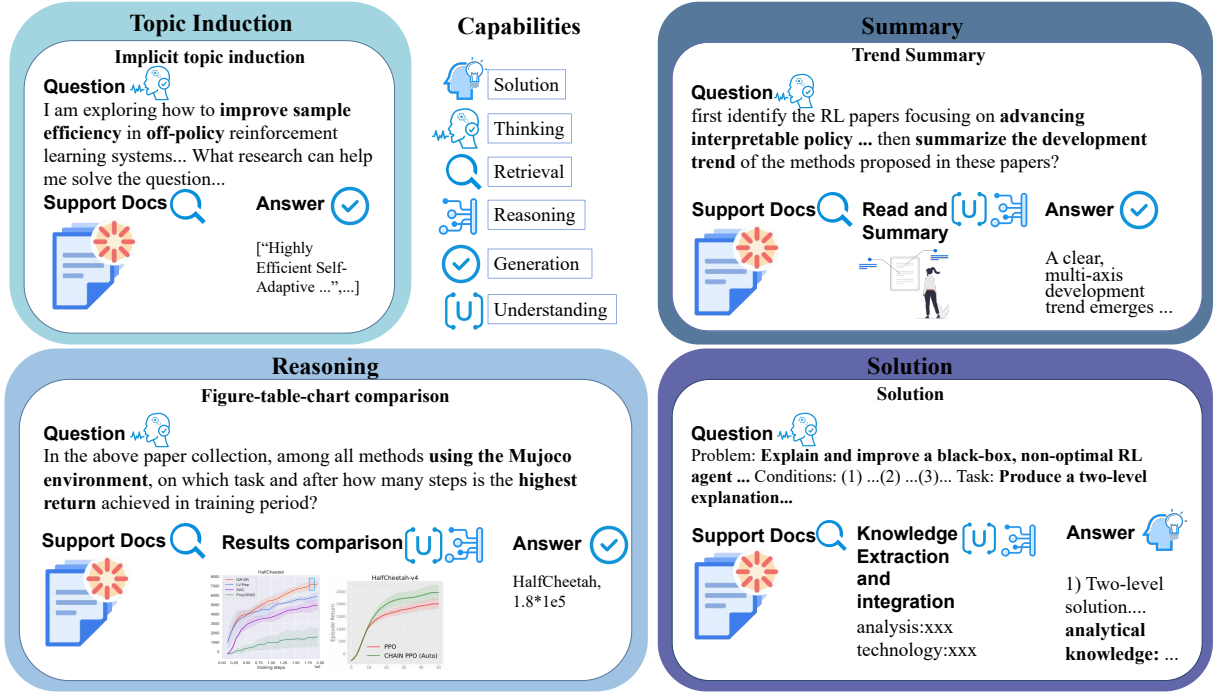


Figure 1: Visualized Examples of PaperScope Bench: Sub-task illustrations from four meta-tasks. The icons in the center represent the various capabilities required by the agent. In each case, the icons placed next to specific stages indicate the particular capabilities needed at that stage. Thinking refers to the reasoning and decomposition of the underlying intent of a given query. Understanding denotes multi-modal comprehension of the document content.

model must generate structured, coherent summaries across three sub-tasks: (1) *Trend Summary*, characterizing research trajectories and emerging directions; (2) *Method Summary*, summarizing strengths and weaknesses of related methods; (3) *Fine-Grained Summary*, comparing experimental results or metrics across studies.

Solution Tasks. Inspired by DeepSolution (Li et al., 2025b), this category evaluates scientific problem solving and innovation. Beyond retrieval and understanding, the model must integrate multi-modal and multi-document evidence to propose actionable, well-grounded solutions. The single sub-task, *Solution Generation*, requires retrieving relevant literature, extracting cross-modal evidence, and producing a comprehensive solution with explicit use of figures, algorithms, and results.

3.4 Statistics of Benchmark

PaperScope comprises a total of 2,400 multi-modal, multi-document questions distributed across four major categories. Specifically, the number of questions for each task and sub-task is summarized in Table 2. Further details on dataset composition, sampling, and annotation procedures are provided in Appendix A and B.

| Category | Sub-Task | Num |
|--------------------------|-------------------------------|-----|
| Multi-Doc Reasoning Task | Figure-table-chart comparison | 200 |
| | Figure-table-chart reasoning | 200 |
| | Formula reasoning | 200 |
| | Algorithm reasoning | 200 |
| | Full paper reasoning | 200 |
| Topic Induction Task | Implicit topic induction | 200 |
| | Explicit topic induction | 200 |
| Summary Task | Trend summary | 200 |
| | Method summary | 200 |
| | Fine-grained summary | 200 |
| Solution Task | Solution generation | 400 |

Table 2: Statistics of task categories and subtasks.

4 Bench Construction Methodology

4.1 Overview

In order to ensure the relevance of the seed papers collection and the difficulty level of the constructed problems, we select the seed paper collection by constructing a knowledge graph of the local large documents and employing random walks. We propose a heterogeneous large-graph-based multi-document association framework for scientific paper selection, designed to ensure answer uniqueness and fully leverage multi-modal evidence in multi-source scenarios. The framework consists of three stages, which is illustrated in Fig-

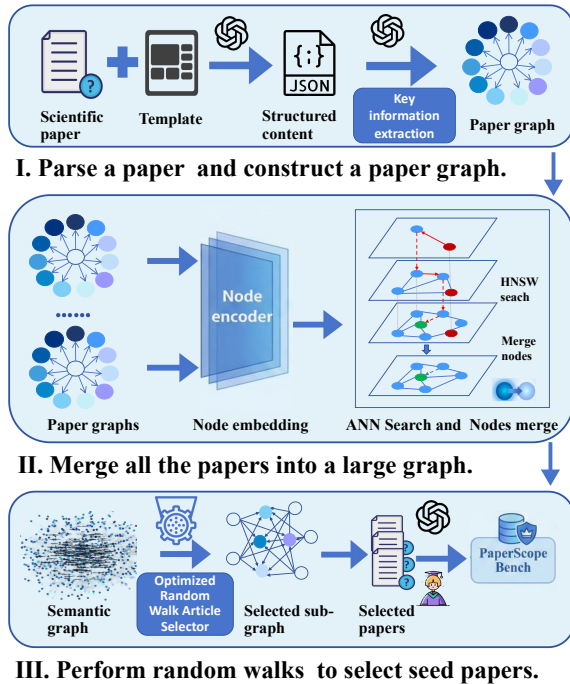


Figure 2: The overview of construction methodology of the PaperScope Bench.

Figure 2: (1) document-level key information extraction and graph construction; (2) efficient nearest-neighbor retrieval and semantic disambiguation merging based on the Hierarchical Navigable Small World (HNSW) graph (Malkov and Yashunin, 2018), which enables logarithmic-time approximate nearest-neighbor search through a multi-layer small-world topology; (3) performing Optimized Random Walk-based Article Selection (ORWAS) on large graphs to identify high-quality papers sharing common key information nodes, which are then used to construct PaperScope.

4.2 Graph Construction

For every paper, we extract structured key information nodes with a large language model, following prior evidence that LLMs perform well on scientific information extraction (Dagdelen et al., 2024). We consider 13 key information nodes: *title*, *research background*, *classification tags*, *key contributions*, *methodology*, *datasets*, *results*, *metrics*, *formulas*, *algorithms*, *figures*, *tables*, and *limitations*. Each paper becomes a title node connected to its key information nodes via “belongs-to” edges.

4.3 Semantic Disambiguation and Merging

To ensure the relevance of the documentation, we merge the knowledge graphs of each paper. Key information nodes across papers are linked when

they are semantically related. All nodes are embedded into a 4096-dimensional space using a shared encoder with type embeddings to accommodate heterogeneous content, including text, figures, tables, algorithms, and formulas. To consolidate key information nodes globally and efficiently, we index all node embeddings with HNSW (Malkov and Yashunin, 2018) and retrieve Top-K neighbors per node. Similarities are computed only within the same key information type, and nodes exceeding a threshold θ are merged while preserving provenance for traceability. We adopt a coarse-to-fine schedule: frequent and clearer types (e.g., classification tags and datasets) are merged first, followed by semantically richer types (e.g., figures, algorithms, and formulas). Typical HNSW settings use $M=32$, $efConstruction=50$, and $efSearch=30$ with $K=20$.

4.4 Optimized Random-Walk Article Selector (ORWAS)

The selection process operates on a heterogeneous graph (approx. 30,000 nodes, 200,000 edges). To ensure efficiency and quality, we employ the following strategies: (1) **Graph Optimization**: Adjacency pre-indexing and neighbor caching reduce access latency to near-constant time. (2) **Sampling & Walk Strategy**: Initialization is stratified (70% article nodes, 30% high-frequency key information nodes). We use a structure-aware random walk: transitions from paper nodes are biased towards high-frequency key information nodes, while others are uniform. (3) **Scalability**: We implement batched parallelization, local accumulation, and strict memory caps on candidate size to prevent overload. (4) **Ranking**: Candidates undergo constrained enumeration and are ranked via a composite score (coverage, diversity, consistency, and redundancy). Details are in Appendix B.

Based on the paper sets and shared key information nodes identified by ORWAS, we construct the benchmark using 4 distinct strategies (details in Appendix B).

4.5 Quality Control

Answer Uniqueness. Adopting BrowseComp’s (Wei et al., 2025b) inverted formulation strategy, we systematically and rationally derive questions from known facts to steadily and reliably converge on unique correct answers, thereby minimizing open-world ambiguity and ensuring strict and robust verifiability.

Multi-modal Dependence. To prevent textual shortcuts, key evidentiary cues are exclusively embedded in visual modalities (figures, tables).

Intrinsic Difficulty. We enforce difficulty via a robustness screen: questions are retained only if SOTA models (Gemini 2.5-Pro (Google, 2025b), GPT-5 (OpenAI, 2025b)) fail to solve them given a single web-search attempt.

5 Experiments

5.1 Evaluation

PaperScope includes 4 task categories that comprehensively evaluate the end-to-end scientific deep research agent from retrieval and understanding to synthesis and problem solving. Each task uses carefully tailored metrics, with detailed formulas and prompts fully provided in Appendix D.

(1) Topic Induction: We report **Recall@K** to measure coverage and topic localization. (2) Reasoning: Performance is measured using **Exact Match (EM)**, reflecting the reliability of multi-source reasoning. (3) Summary: We use a hybrid evaluation combining induction scores with a GPT-5 score across five dimensions: **fluency, relevance, accuracy, creativity, and overall quality**, following ResearchPulse (Chen et al., 2025). (4) Solution: Following SolutionBench (Li et al., 2025b), we use **Analysis and Technical Scores** to assess problem decomposition, reasoning design, and domain-knowledge usage, with scores judged by GPT-5 (OpenAI, 2025b).

5.2 Settings and Baselines

5.2.1 Experiment Settings

We evaluate agents in a unified multi-modal retrieval and reasoning environment for multi-document scientific tasks. Inputs include high-resolution images and structurally parsed PDF–Markdown documents (processed with MinerU2.5 (Niu et al., 2025)), enabling joint text–image reasoning over figures, tables, formulas, and pseudocode.

Models follow the ReAct (Yao et al., 2022) paradigm, alternating between reasoning and action. They plan via chain-of-thought, invoke tools for retrieval and parsing, and generate final answers. All models use two tools: (i) **Local FileSearch** for semantic retrieval using an Ops-MM embedding model; (ii) **Local FileVisit** for accessing relevant files and returning high-resolution images and structured PDF–Markdown content.

We construct a stratified subset of 200 QAs (50% reasoning, 30% summary, 10% induction, 10% solution) guided by three principles: (1) **Objective evaluation** prioritizing verifiable logical structures; (2) **Multi-modal rigor** requiring integration of complex cross-document elements (e.g., figures); and (3) **Pipeline completeness** encompassing the full retrieval-to-synthesis workflow.

5.2.2 Selected Baselines

We evaluate baselines across two categories: (1) **MLLM-based ReAct agents**, including WebWatcher 32B (Geng et al., 2025), kimi k2 (Team et al., 2025a), GLM 4.5V (Team et al., 2025d), Qwen3-VL (Qwen team, 2025), DeepSeek-V3.1 (DeepSeek-AI, 2024), OpenAI 4o-mini (OpenAI, 2024), Gemini-2.5-flash-thinking, Gemini-2.5-pro (Google, 2025b), and OpenAI GPT-5.1; and (2) **Deep Research-specialized models**, such as DR Tulu-8B (Shao et al., 2025), MMSearch-R1-7B (Wu et al., 2025), Asearcher-Web-7B (Gao et al., 2025), MiroThinker-v1.0-30B (Team et al., 2025b), Tongyi Deep Research 32B (Team et al., 2025c), OpenAI o3 Deep Research (OpenAI, 2025a), and Grok 4 (xAI, 2025). See more baseline details in Appendix C.

5.3 Main Results

For each task, macro-averaged scores are reported in Table 3, with more detailed results provided in Appendix D. In addition, we present case studies of different model categories to illustrate systematic performance variations across task types, as shown in Figure 4. Based on these results, several key observations can be drawn:

(1) **Deep-research models show advantages in reasoning and tool-mediated integration.** In the heavily weighted Reasoning task, deep-research agents outperform LLM-based ReAct baselines, reflecting stronger capacity for cross-document and cross-modal evidence integration involving tables, figures, formulas, and algorithms. Representative models such as Grok-4 and OpenAI o3 deep research achieve comparatively high scores across Reasoning, Summarization, and Solution tasks. This advantage depends critically on stable tool-calling. However, several deep-research models and MLLMs fail to execute tools reliably, which substantially degrades their reasoning and induction performance.

(2) **ReAct models remain effective for summarization but fall short in scientific problem**

| Agent Type | Model | Reasoning | Induction | Summary | Solution | Score |
|-----------------|---------------------------|-----------|--------------|--------------|--------------|--------------|
| LLM-based ReAct | WebWatcher 32B | 4 | 0 | 46.74 | 26.78 | 18.70 |
| | OpenAI 4o-mini | 6 | 25.49 | 53.26 | 22.1 | 23.74 |
| | Gemini-2.5-flash-thinking | 7 | 13.33 | 38.40 | 29.71 | 19.32 |
| | OpenAI GPT-5.1 | 0 | 0 | 42 | 51.84 | 17.78 |
| | Gemini 2.5 pro | 3 | 7.02 | 47.54 | 40.39 | 20.50 |
| | GLM 4.5V | 0 | 0 | 37.32 | 32.45 | 14.44 |
| | Kimi k2 | 12 | 24.07 | 56.64 | 49.85 | 30.38 |
| | Qwen3-VL | 4 | 13.33 | 52.74 | 37.38 | 22.89 |
| | deepseek-V3.1 | 6 | 26.32 | 52.22 | 51.65 | 26.46 |
| Deep Research | DR Tulu-8B | 4 | 0 | 40.60 | 38.71 | 18.05 |
| | MMSearch-R1-7B | 8 | 3.70 | 43.66 | 17.21 | 19.19 |
| | ASearcher-Web-7B | 13 | 0 | 47.26 | 8.95 | 21.57 |
| | MiroThinker-v1.0-30B | 3 | 3.92 | 27.44 | 32.03 | 13.33 |
| | Tongyi Deep Research 30B | 11 | 0 | 5 | 36.55 | 10.66 |
| | OpenAI o3 deep research | 13 | 0 | 56.26 | 59.15 | 29.29 |
| | Grok 4 | 36 | 20 | 53.74 | 48.28 | 40.95 |

Table 3: Performance of different agents across Reasoning, Induction, Summary, and Solution tasks. The best score on each dataset is shown in bold.

PaperScope Bench Ablation Analysis

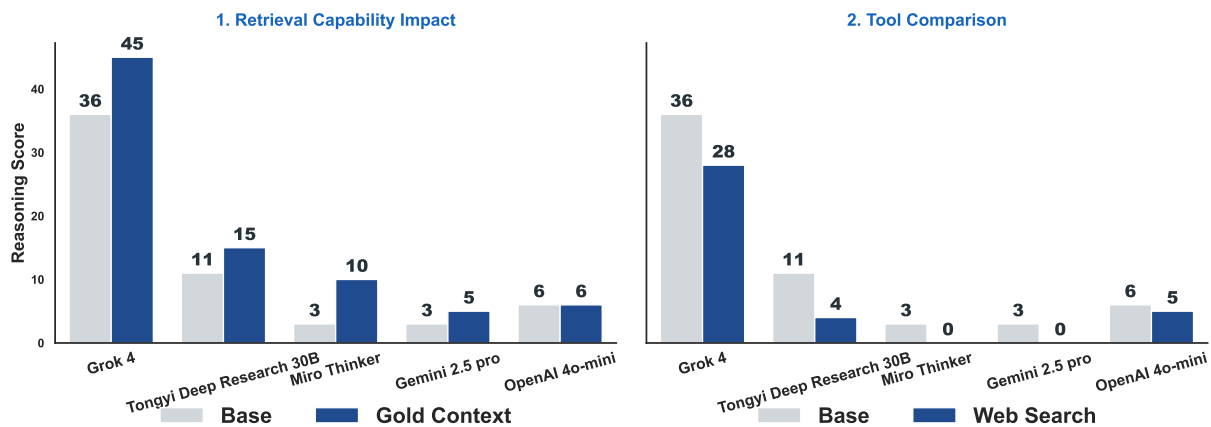


Figure 3: The ablation results of PaperScope Bench in reasoning task.

solving. As shown in Table 3, ReAct-based systems such as Kimi k2, Qwen3-VL, and DeepSeek-V3.1 achieve strong Summarization results. Their overall performance is nevertheless constrained by weak Reasoning and Solution scores, indicating limited ability to perform multi-source synthesis, cross-modal alignment, and structured inference required in scientific settings.

(3) **Complex multi-modal reasoning constitutes a universal bottleneck.** Most models struggle to interpret figures, tables, and mathematical expressions, leading to consistently low Reasoning scores. Observed failures include broken logical chains, misaligned evidence, and insufficient multi-modal fusion. The sensitivity of some deep-

research models to tool usage further indicates overfitting in their tool-calling strategies, which partially explains the lower performance of Tongyi Deep Research 30B. These results highlight persistent challenges in visual-semantic reasoning, evidence chain construction, and systematic decomposition.

(4) **Long-context modeling limits performance across paradigms.** Across paradigms, Reasoning performance remains low, indicating difficulty in inferring research intent and managing heterogeneous inputs. Performance plateaus below 60 scores even in Summarization and Solution tasks, suggesting a reliance on parametric knowledge over reasoning. Extended multi-turn interac-

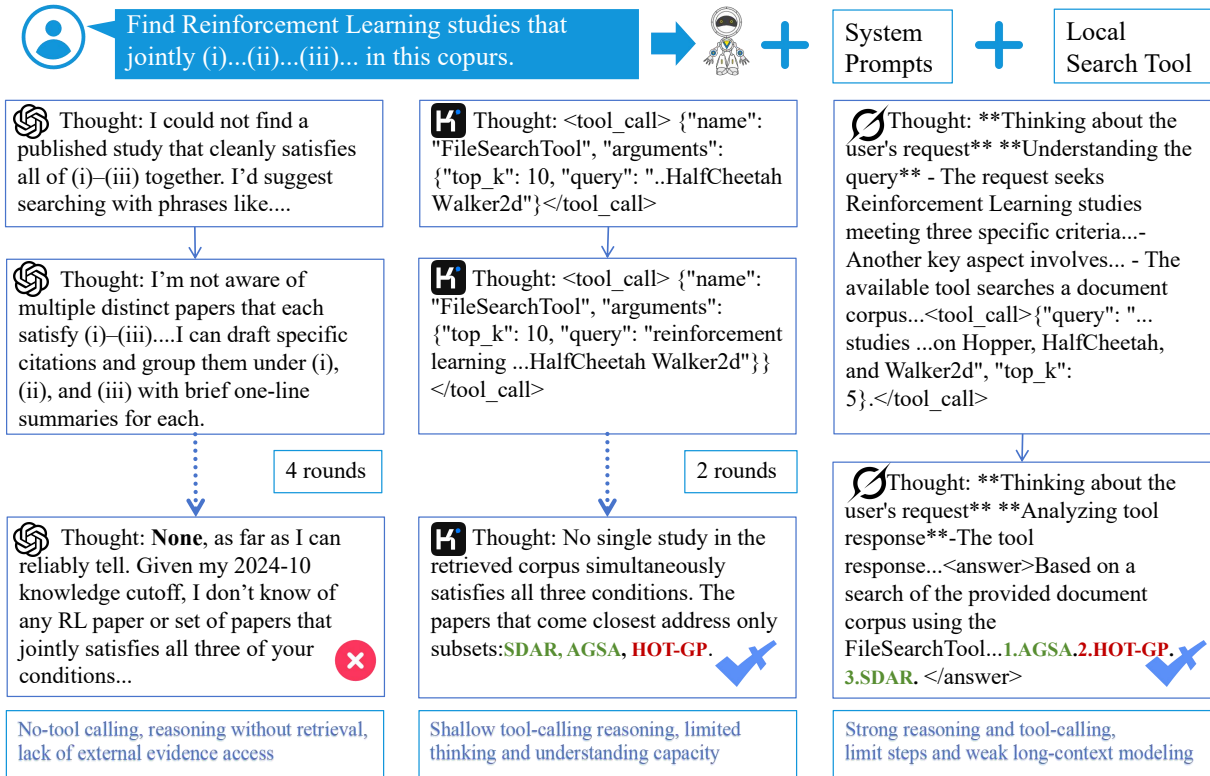


Figure 4: A case study comparing the capabilities of different model types in tool use and reasoning. The figure illustrates the behavioral distinctions and performance gaps between models with no tool support (left), shallow tool calling (middle), and strong reasoning combined with tool calling (right) when addressing complex retrieval tasks. For brevity, intermediate tool return results and the FileVisit tool interactions are omitted.

tions frequently trigger catastrophic forgetting and hallucinations, which, combined with unreliable tool invocation, destabilize reasoning trajectories and constrain practical deployment.

5.4 More Ablation Analysis

(1) **Retrieval quality matters, but reasoning is the dominant bottleneck.** Directly providing oracle support documents improves performance (Figure 3 (1)), yet neither open-source nor closed-source models achieve strong results. This indicates that while retrieval is a factor, the primary bottleneck lies in the inherent complexity of the cross-source multi-modal reasoning required by the benchmark. More ablation experiment details are shown in Appendix D.

(2) **Domain-specific local retrieval outperforms generic WebSearch.** To examine the impact of the provided local retrieval tool on model performance, we replace it with a commonly used community web search tool. As shown in Figure.3 (2), the evaluation scores of the tested models drop substantially. This decline can be attributed to the models’ inability to reliably retrieve the cor-

rect support documents when using web search, causing them to rely predominantly on knowledge encoded in their model parameters rather than evidence grounded in retrieved documents.

(3) **Optimal ORWAS hyperparameters balance combination diversity and computational cost.** We conducted a grid search on `corpus_test` to evaluate the core hyperparameters of the ORWAS algorithm: Walk Length (L), Bias Probability (β), and Number of Walks (W). To measure effectiveness, we introduce a *Combination Quality Score*—a weighted metric based on shared entity importance and coverage. A higher score implies tighter semantic coupling between papers, enabling human annotators to construct harder, more logical reasoning chains.

As shown in Table 4, our default settings ($L = 100, \beta = 0.3, W = 10000$) maximize diversity (measured by unique combinations) at a low computational cost. Specifically, short walk lengths ($L = 10$) fail to escape local nodes, while excessively long walks ($L \geq 500$) introduce noise or diverge (Table 4 (a)). For bias probability, heavy exploitation ($\beta = 1.0$) causes paths to collapse onto

| Setting | Combinations (Uniqueness) | Max Quality Score | Interpretation |
|--|---------------------------|-------------------|--|
| <i>(a) Walk Length (L) — finding the semantic sweet spot (with $W = 10000, \beta = 0.3$)</i> | | | |
| $L = 10$ | 2 | 60.33 | Too Short: Fails to escape local nodes. |
| $L = 100$ (Default) | 10,000 (Max) | 1,564.19 | Optimal: Best balance of diversity and efficiency. |
| $L = 500$ | 10,000 (Max) | 26,430.99 | Diminishing Returns: Deep links but introduces noise. |
| $L = 1000$ | - | - | Divergence: Paths drift too far, failing to converge. |
| <i>(b) Bias Probability (β) — exploration vs. exploitation (with $L = 100, W = 10000$)</i> | | | |
| $\beta = 0.3$ (Default) | 10,000 (Max) | 1,564.19 | Balanced: Effectively explores long-tail entities. |
| $\beta = 0.7$ | 10,000 (Max) | 1,789.55 | Exploitation: Favors hubs, slightly higher scores. |
| $\beta = 1.0$ | 800 | 1,693.88 | Collapse: Fixates on hubs, destroying diversity. |
| Setting | Combinations (Uniqueness) | Time Cost | Interpretation |
| <i>(c) Number of Walks (W) — coverage vs. cost (with $L = 100, \beta = 0.3$)</i> | | | |
| $W = 1000$ | 16 | < 10s | Under-sampling: Sparse trajectories miss targets. |
| $W = 10,000$ (Default) | 10,000 (Max) | ~ 60s | Efficiency Saturation: Captures most targets quickly. |
| $W = 50,000$ | 10,000 (Max) | ~ 480s | High Cost: 8× slower for negligible diversity gain. |

Table 4: Ablation analysis of ORWAS hyperparameters. Parameter tuning significantly impacts the uniqueness (discovery rate) of document combinations. Our default settings yield the maximum combination diversity without incurring excessive computational overhead.

high-frequency hubs, drastically reducing uniqueness (Table 4 (b)). Finally, $W = 10,000$ reaches efficiency saturation in roughly 60 seconds; further scaling incurs higher time costs without significant diversity gains (Table 4 (c)). It is important to note that while ORWAS determines the *structural complexity* (i.e., the logic chain), the final *solving difficulty* is primarily driven by the multi-modal information density (e.g., charts and tables) within the retrieved papers.

5.5 Cross-Domain Generalizability

To verify broader applicability, we conducted a pilot study on 24 papers from Medicine and Mechanics. By minimally adapting the node schema to domain-specific structures (e.g., adding *Limitations* for Medicine and *Equations* for Mechanics), our pipeline constructed high-density knowledge graphs (> 690 nodes and > 35,000 edges per domain) achieving a 95% expert satisfaction rate. Furthermore, the framework successfully generated domain-specific multi-modal reasoning tasks, such as joint-chart analyses for clinical effect sizes and cross-verifying tabular data with theoretical mechanics formulas. Crucially, our automated

random-walk algorithm seamlessly linked these cross-document entities and modalities without requiring domain-specific heuristic tuning. These results demonstrate our method’s strong cross-domain transferability with adaptation costs.

6 Conclusion

In this paper, we present PaperScope, a comprehensive multi-modal, multi-document benchmark tailored for the challenges of scientific deep research. Constructed from knowledge graphs spanning over 2,000 AI papers, PaperScope synthesizes semantically aligned information nodes and employs random walk algorithms to sample seed papers. This process yields over 2,000 high-quality QA pairs encompassing 4 distinct capabilities: topic induction, reasoning, summarization, and solution generation. Experimental evaluations reveal that even state-of-the-art agents face significant challenges (Grok 4’s score of 40.95), underscoring the persistent difficulties in long-context retrieval and multi-document reasoning. Beyond serving as an evaluation suite and a scalable construction pipeline, we envision PaperScope as a pivotal resource for model training.

Acknowledgment

This work was supported by National Natural Science Foundation of China No. 62272467. The work was partially done at the Beijing Key Laboratory of Research on Large Models and Intelligent Governance and Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

Limitations

PaperScope is constructed on a large but finite local corpus rather than a fully open-domain setting, which may not fully reflect challenges arising from unbounded document collections and noisy retrieval. Moreover, ORWAS adopts heuristic design choices, such as biased random walks and constrained combination enumeration, which improve efficiency but do not guarantee globally optimal article selection. In addition, query and annotation generation rely on large language models with prompt-based control, where residual noise and model bias may affect evaluation stability. Finally, due to computational constraints, experiments are conducted with limited baseline coverage and fixed inference budgets, leaving more extensive evaluations to future work.

References

- Qi Chen, Jingxuan Wei, Zhuoya Yao, Haiguang Wang, Gaowei Wu, Bihui Yu, Siyuan Li, and Cheng Tan. 2025. Researchpulse: Building method-experiment chains through multi-document scientific inference. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9110–9119.
- Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, and 1 others. 2025. Curie: Evaluating llms on multitask scientific long-context understanding and reasoning. In *The Thirteenth International Conference on Learning Representations*.
- John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.
- DeepSeek-AI. 2024. *Deepseek-v3 technical report. Preprint*, arXiv:2412.19437.
- Kuicai Dong, Shurui Huang, Fangda Ye, Wei Han, Zhi Zhang, Dexun Li, Wenjun Li, Qu Yang, Gang Wang, Yichao Wang, and 1 others. 2025. Doc-researcher: A unified system for multimodal document parsing and deep research. *arXiv preprint arXiv:2510.21603*.
- Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025. *Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. Preprint*, arXiv:2508.07976.
- Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.
- Google. 2025a. *Deep research is now available on gemini 2.5 pro experimental*.
- Google. 2025b. Gemini 2.5 pro. <https://deepmind.google/technologies/gemini/pro/>.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.
- Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024a. M3sciq: A multi-modal multi-document scientific qa benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multi-modal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387.
- Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025a. Webthinker: Empowering large reasoning models with deep research capability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyong Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024c. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*.
- Zhuoqun Li, Haiyang Yu, Xuanang Chen, Hongyu Lin, Yaojie Lu, Fei Huang, Xianpei Han, Yongbin Li, and Le Sun. 2025b. Deepsolution: Boosting complex engineering solution design via tree-based exploration and bi-point thinking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4396.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc:

- Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.
- Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023a. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023b. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, and 1 others. 2025. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*.
- OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- OpenAI. 2025a. [Deep research system card](#).
- OpenAI. 2025b. Gpt-5 is here. <https://openai.com/gpt-5/>.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*.
- Alibaba Cloud Qwen team. 2025. [Qwen3-vl: Sharper vision, deeper thought, broader action](#).
- Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, and 1 others. 2025. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025a. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025b. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.
- Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025c. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.
- V Team, Wenyi Hong, Wenmeng Yu, and 1 others. 2025d. [Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Yang Tian, Zheng Lu, Mingqi Gao, Zheng Liu, and Bo Zhao. 2025. Mmcr: Benchmarking cross-source reasoning in scientific papers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 488–497.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.
- Haoran Wei, Yaofeng Sun, and Yukun Li. 2025a. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.
- Jason Wei, Zhiqing Sun, Spencer Papay, and 1 others. 2025b. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.
- Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025. Mmsearch-r1: Incentivizing llms to search. *arXiv preprint arXiv:2506.20670*.
- xAI. 2025. [Grok 4](#).
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

A Detailed Overview of PaperScope

A.1 Detailed Benchmark Statistic

PaperScope consists of multiple large-scale corpus constructed from recent AI conference papers. To facilitate reproducibility, we report detailed corpus-level statistics in Table 5.

| Corpus Split | #Documents | Source Venues | #QA Pairs |
|--------------|------------|---------------|-----------|
| Corpus-A | 500 | 5 conferences | 1100 |
| Corpus-B | 1500 | 5 conferences | 1100 |
| Test Corpus | 202 | 5 conferences | 200 |

Table 5: High-level statistics of PaperScope corpus.

We further analyze the distribution of answer-support documents per question in Table 6, which reflects the degree of multi-document dependency required by each task.

| 1 Doc | 2 Docs | 3 Docs | 4 Docs | 5+ Docs |
|-------|--------|--------|--------|---------|
| 200 | 200 | 400 | 200 | 1400 |

Table 6: Distribution of the number of supporting documents per question.

A.2 Data Collection

All documents utilized in this study are strictly sourced from repositories governed by the Creative Commons 4.0 (CC 4.0) licensing framework. Furthermore, the constructed queries were designed to focus exclusively on content analysis, ensuring the complete exclusion of personally identifiable information (PII) or sensitive private data regarding the authors.

B More Detailed of Data Construction

B.1 Detailed Graph Statistic

We construct two large heterogeneous semantic graphs from different corpora. Table 7 summarizes their structural properties. Figure 5 visualizes the graphs.

| Statistic | Graph-Test | Graph-Docs 500 | Graph-Docs 1500 |
|------------------|------------|----------------|-----------------|
| #Nodes | 30962 | 31348 | 85152 |
| #Edges | 1990449 | 1907081 | 12735443 |
| Density | 0.0042 | 0.0039 | 0.0025 |
| Diameter | 3 | 4 | 8 |
| Avg. Degree | 128.57 | 121.67 | 125.81 |
| Avg. Path Length | 2.041 | 2.17 | 4.14 |

Table 7: Structural statistics of the constructed semantic graphs.

Algorithm 1 Stratified Random Walk Sampling

Input: Heterogeneous graph $G = (V, E)$; Article nodes V_a ; Key information nodes V_e ; High-frequency key information nodes V_h ; Walk length L ; Number of walks W ; Bias probability β

Output: Article-key information node map \mathcal{M}

```
1: Initialize neighbor cache  $\mathcal{N}(v)$  for all  $v \in V$ 
2: Initialize article-key information node map  $\mathcal{M} \leftarrow \emptyset$ 
3: Sample  $0.7W$  start nodes from  $V_a$  and  $0.3W$  from  $V_h$ 
4: for all start node  $s$  in parallel do
5:    $path \leftarrow [s]$ 
6:   for  $i = 1$  to  $L$  do
7:      $u \leftarrow$  last node in  $path$ 
8:     if  $u \in V_a$  and  $\text{rand}() < \beta$  then
9:       Select  $v$  uniformly from  $\mathcal{N}(u) \cap V_h$ 
10:      if empty then
11:        Select  $v$  uniformly from  $\mathcal{N}(u)$ 
12:      end if
13:    else
14:      Select  $v$  uniformly from  $\mathcal{N}(u)$ 
15:    end if
16:    Append  $v$  to  $path$ 
17:  end for
18:  Extract article-key information node pairs from  $path$  and update  $\mathcal{M}$ 
19: end for
20: return  $\mathcal{M}$ 
```

B.2 Detailed Optimized Random Walk Article Selector (ORWAS)

ORWAS operates on a heterogeneous paper-key information node graph to identify compact yet thematically coherent paper subsets. Detailed core algorithm is shown in Algorithm 1. Table 8 summarizes the time and space complexity.

| Stage | Time Complexity | Space Complexity |
|--------------------|-------------------------|------------------|
| Random Walk | $O(W \cdot L)$ | $O(L)$ |
| Aggregation | $O(A \cdot E)$ | $O(A)$ |
| Combination Filter | $\min(C, \binom{N}{k})$ | Linear in output |

Table 8: Complexity analysis of ORWAS.

B.3 Detailed Task Formulation

Detailed Tasks' Prompt For the scientific reproducibility, we provide prompts for 3 types of synthetic tasks. Among them, the summary-type tasks include a dedicated prompt for trend questions, as shown in the Figure 6, 7, and 8.

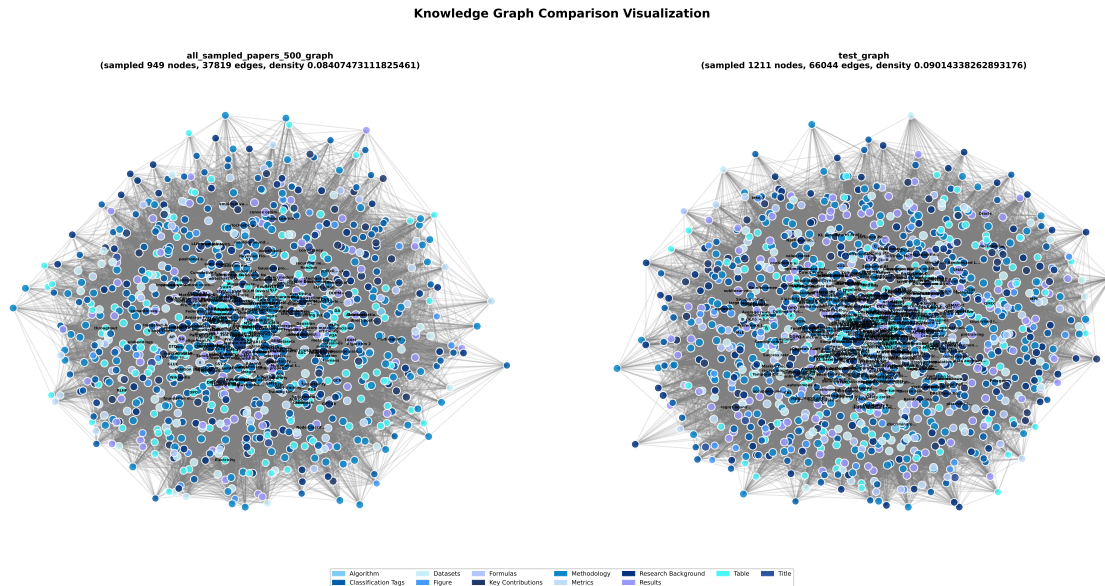


Figure 5: Visualization of selected semantic graphs.

B.4 Detailed Annotations

To ensure the PaperScope benchmark maintains high intrinsic difficulty and strictly necessitates multi-modal reasoning, we implemented a bifurcated data construction strategy. This process distinguishes between expert annotation for complex reasoning tasks and a rigorous human-in-the-loop verification protocol for structured extraction tasks (Topic Induction, Summary, and Solution).

B.4.1 Expert Annotation for Cross-Document Reasoning

For the *Reasoning* task category, computer science graduate students were engaged to formulate complex queries based on document clusters identified by the Optimized Random-Walk Article Selector (ORWAS). The annotation workflow was governed by an “inverted construction” paradigm designed to eliminate open-world ambiguity. Annotators first isolated irrefutable evidence nodes (ground truth) across multiple documents before deriving the corresponding questions, thereby guaranteeing *Answer Uniqueness* and a closed retrieval scope.

A critical constraint imposed during this phase was *Multi-modal Dependence*. Annotators were instructed to select evidence exclusively from visual modalities such as trend lines in ablation studies, numerical entries in comparative tables, or architectural connections in model diagrams. Questions were deemed valid only if they required cross-referencing these visual anchors with textual context; items solvable via text-only shortcuts were

systematically rejected. Furthermore, to enforce *Intrinsic Difficulty*, the protocol required that questions necessitate multi-hop inference. Queries resolvable through lexical matching or single-step web retrieval were discarded, ensuring the benchmark challenges the reasoning upper bounds of current state-of-the-art models.

B.4.2 Expert Verification for Structured Tasks

For *Topic Induction*, *Summary*, and *Solution* tasks, we adopted an expert auditing mechanism to validate the quality of semi-automated candidates.

- **Topic Induction Audit:** Reviewers verified that aggregated multi-modal nodes formed a coherent thematic query. The core criterion was exclusivity: the selected paper cluster must serve as the unique and optimal ground truth for the induced topic, ruling out weak or tangential associations.
- **Summary and Solution Audit:** For *Summary* tasks, experts scrutinized the integration of fine-grained experimental comparisons with coarse-grained methodological trends, specifically checking for accurate temporal alignment and logical chart merging. In the *Solution* category, validation focused on the structural integrity of the extracted knowledge, ensuring that specific conditions and challenges were logically organized under the unified theme without hallucination.

B.5 Human-Model Evaluation Alignment

To demonstrate the reliability of our automated evaluation framework (particularly the GPT-5 judge), we conducted a blind Human-Model Alignment Study. We randomly sampled 10 instances each from the *Summary* and *Solution* tasks across 6 different models, yielding a total of 120 evaluation trajectories. Three CS Ph.D. students independently scored these trajectories using a streamlined rubric.

As detailed in Table 9, there is a strong and statistically significant positive correlation ($p < 0.0001$) between the automated evaluator and human judgments. According to established literature, a correlation coefficient surpassing 0.6 indicates substantial consensus. Notably, the automated judge performs exceptionally well on the reasoning-intensive *Solution* task (Pearson $r = 0.6385$). Specifically, the correlation for the "Analysis" dimension peaked at 0.7173, indicating that the evaluator aligns closely with experts in identifying complex logical steps. Furthermore, its high Spearman correlation ($\rho > 0.6$) confirms strong ranking stability, validating its reliability as a diagnostic metric for agentic systems.

| Task Type | Pearson (r) | Spearman (ρ) | P-value |
|-----------|-----------------|---------------------|------------|
| Summary | 0.6175 | 0.6072 | < 0.0001 |
| Solution | 0.6385 | 0.6263 | < 0.0001 |

Table 9: Correlation between automated evaluation and human expert judgments based on 120 trajectories.

B.6 Multi-modal Dependency and Data Verification

Multi-modal Dependency. To verify whether models genuinely engage in multi-modal reasoning rather than relying on textual shortcuts, we conducted a systematic Text-only Ablation experiment on a subset of 100 reasoning tasks. Visual inputs (e.g., charts and figures) were removed, leaving only OCR-extracted text. As shown in Table 10, performance drops precipitously—by an average of 81.7%—when visual inputs are absent. Only approximately 8% of the questions could be solved via pure text (typically when specific values were explicitly described in the main text). This heavy visual dependency is driven by our "Visual Anchor" design strategy, which explicitly requires operations like cross-comparing a table’s data with a figure’s curve. Without visual perception, the

logical chain inherently breaks.

| Model | Full Score | Text-only | Drop (%) |
|----------------------|------------|-----------|---------------|
| WebWatcher 32B | 4 | 1 | 75.0% |
| Gemini 2.5 pro | 3 | 1 | 66.7% |
| Kimi k2 | 12 | 3 | 75.0% |
| MiroThinker-v1.0-30B | 3 | 0 | 100.0% |
| OpenAI o3 (Deep Res) | 13 | 3 | 76.9% |
| Grok 4 | 36 | 5 | 86.1% |

Table 10: Text-only ablation results. The massive performance drop indicates that the benchmark heavily relies on visual information rather than simple text matching.

Answer Uniqueness and Verification Statistics. In our benchmark, "uniqueness" refers to constraint-based determinability within the local corpus. To guarantee data quality, we implemented a rigorous dual-verification mechanism on Corpus-A:

- **Stage 1: AI-based Difficulty Filtering.** We used an ensemble (Gemini-2.5-pro + GPT-5 + Web Search) to filter out trivial instances solvable via direct retrieval. This removed 9.98% of the samples (112 items).
- **Stage 2: Human Validity Verification.** Expert annotators performed a secondary review to check for exclusivity, semantic coherence, and structural completeness. This filtered an additional 7.45% (82 items), ensuring high reliability and unambiguous reasoning bounds for the retained tasks.

C More Detailed of baseline

All experiments are conducted on a multi-GPU environment. Table 11 summarizes the default configuration.

| Component | Configuration |
|------------------|------------------------|
| GPUs | $8 \times \text{H100}$ |
| Inference Engine | vLLM |
| Max retries | 10 |
| Max excute time | 2h30mins |
| Max tokens | 128000 |
| Top p | 0.95 |
| Temperature | 0.6 |
| Batch Size | 4 |
| presence penalty | 1.1 |

Table 11: Experimental configuration for baseline evaluation.

D Detailed Evaluation of PaperScope

D.1 Evaluation Prompts

The specific evaluation prompts utilized for the Summarization and Solution tasks are delineated in Figure 9, Figure 10 and 11.

For the **Summary task**, performance is assessed across five distinct dimensions: *Fluency*, *Relevance*, *Accuracy*, *Creativity*, and *Overall Quality*. In this process, GPT-5 serves as the adjudicator, evaluating the efficacy of system-generated responses against gold-standard references based on these metrics.

Regarding the **Solution task**, evaluation is quantified via two primary metrics: the *Analysis Score* and the *Technology Score*. As illustrated in the Figure 10 and 11, we synthesized tailored evaluation rubrics for each component. To ensure robust assessment, the evaluator is supplied with a comprehensive context window comprising relevant analysis knowledge, technical knowledge, golden explanation and golden solution extracted from the source text.

D.2 Detailed Tools

The proposed FileSearchTool module is designed to construct an efficient document retrieval system based on semantic similarity, capable of processing a heterogeneous corpus containing Markdown text and various image formats (e.g., PNG, JPG). The core logic is encapsulated within the FileSearchEngine class, which orchestrates multi-modal embedding and Optical Character Recognition (OCR) technologies to achieve unified indexing and retrieval across modalities.

In terms of model architecture, the system integrates two pivotal pre-trained models. First, the Ops-MM-embedding-v1³ model is employed as a unified feature encoder, ensuring that text and image inputs possess commensurability within the same high-dimensional space. Second, the DeepSeek-OCR (Wei et al., 2025a) model is introduced for deep semantic understanding and text extraction. To optimize efficiency, the system utilizes the FAISS library to construct an IndexFlatIP index based on inner product similarity. During index construction, text content is extracted directly, while visual information is transformed into vector

³Ops-MM-embedding-v1-7B:
<https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-7B>

representations and stored alongside text embeddings to support unified retrieval.

In the retrieval phase, user queries are encoded to identify the top- k nearest candidates. The system applies modality-aware post-processing: text documents are returned directly, while images undergo OCR and captioning via the prompt “<image>\nCaption this image.” to generate textual descriptions. Integrated into the Qwen-Agent framework as a BaseTool, the module processes JSON parameters and returns aggregated content to serve as context augmentation for downstream Large Language Model tasks.

D.3 Detailed Ablation Experiments

To rigorously decouple the contribution of the retrieval module from the reasoning capabilities of the agent, and to assess the benchmark’s resilience against generic information retrieval methods, we conducted a series of ablation experiments. These experiments were exclusively performed on the *Reasoning* task subset. This specific subset was selected as the experimental substrate due to its strict multi-modal dependencies and high intrinsic difficulty, combined with the objectivity of its ground truth, which allows for unambiguous performance evaluation.

D.3.1 Impact of Retrieval Necessity: The Oracle Setting

To investigate whether the agent’s performance is bottlenecked by retrieval accuracy or reasoning limitations, we implemented an experimental setup that bypasses the active search process. In this configuration, the standard retrieval module was replaced by a deterministic mechanism, referred to as the *Direct Context Tool*. This tool was exposed to the model with the following functional description:

“Directly retrieve pre-defined documents based on the original question without any retrieval process. Returns the parsed markdown documents corresponding to the question.”

By feeding the model the ground-truth parsed markdown documents directly, this setup simulates an “oracle” retrieval scenario. Comparing the performance of this configuration against the full pipeline allows us to quantify the gap between the model’s reasoning potential given ideal context and its ac-

tual performance when burdened with the noise and uncertainty of retrieval.

D.3.2 Impact of Generic Web Search

To validate the hypothesis that the PaperScope benchmark requires specialized, domain-specific indexing rather than general open-web knowledge, we evaluated an agent configuration equipped with a commercial-grade search engine. We integrated the Bocha API to construct a *Web Search Tool*, which was defined for the agent as follows:

“Performs batched web searches: supply an array ‘query’; the tool retrieves the top 10 results for each query in one call.”

This ablation assesses the vulnerability of the benchmark to existing commercial search solutions. Under the assumption that the “Answer Uniqueness” and “Multi-modal Dependence” principles (Section 4.5) were effectively implemented, the generic web search is expected to underperform, as the specific visual evidence and cross-document inference paths required for the *Reasoning* tasks are unlikely to be indexed or synthesized effectively by standard search engines.

D.4 Detail Evaluation Results

To provide a more fine-grained characterization of model behavior across different task types, we report detailed results for the induction and solution tasks in which the models achieve non-zero performance, as shown in Table 12 and Table 13. These task-specific analyses are intended to expose systematic performance patterns that are otherwise obscured by aggregate metrics. Due to space constraints, only partial results covering all subtasks are presented for a subset of models, as shown in Table 14.

D.5 Case Study

We present one successful case and one failure case to illustrate typical agent behaviors.

Successful Case. Figure 12 presents a correct case which grok-4 successfully answer the full paper reasoning question.

Failure Case. Figure 13 and Figure 14 show a wrong case which grok-4 didn’t retrieval the all golden documents and understanding all results in the tables and charts.

E Error Attribution Analysis

To understand exactly where agentic pipelines fail on complex academic tasks, we conducted a fine-grained Error Attribution Analysis on the strongest baseline model (Grok-4). Human experts manually reviewed 40 randomly sampled failure cases against the gold-standard documents. The breakdown of failure modes is presented in Table 15.

Based on this analysis, we identify three major bottlenecks in current systems:

1. **multi-modal Information Extraction is the Primary Weakness (37.5%):** Surprisingly, fine-grained visual understanding bottlenecks the pipeline earlier than reasoning. Even when the correct paper is retrieved, the model often “reads but does not understand” (e.g., failing to align rows/columns in complex tables or extracting inaccurate coordinates from line charts).
2. **Error Accumulation in Multi-step Reasoning (22.5%):** In cross-paper synthesis tasks, models frequently err in intermediate steps. Small deviations in initial metric extraction compound during comparisons, causing final answers to drift significantly.
3. **Retrieval Granularity and Hallucination (35% combined):** Approximately 20% of errors stem from broad semantic search scopes failing to pinpoint specific papers. The 15% hallucination rate is often a secondary effect—when exact evidence is missed, models tend to generate speculative answers rather than abstaining.

F Usage of LLM

In the preparation of this manuscript, Large Language Models (LLMs) were utilized to facilitate translation and linguistic refinement. regarding the computational implementation, while LLMs assisted with auxiliary coding routines, the development of critical data processing pipelines and core algorithmic architectures was conducted exclusively by the authors to ensure scientific integrity.

| Model | Max Score | Min Score | Macro Avg. |
|---------------------------|-----------|-----------|------------|
| OpenAI o3 deep research | 79 | 35 | 59.15 |
| kimi-k2 | 73 | 11 | 49.85 |
| Grok-4 | 65 | 32 | 48.27 |
| Gemini-2.5-pro | 51.5 | 19 | 40.39 |
| DR Tulu-8B | 51 | 0 | 38.71 |
| Qwen3-VL | 59 | 25 | 37.38 |
| Tongyi-DeepResearch-30B | 65 | 14 | 36.55 |
| GLM 4.5V | 52 | 9 | 32.45 |
| MiroThinker-v1.0-30B | 100 | 0 | 32.02 |
| Gemini-2.5-flash-thinking | 62 | 0 | 29.70 |
| WebWatcher-32B | 48 | 0 | 26.77 |
| OpenAI 4o-mini | 42.5 | 9 | 22.10 |
| MMSearch-R1-7B | 98 | 0 | 17.21 |
| deepseek-V3.1 | 64.5 | 0 | 14.18 |
| OpenAI GPT-5.1 | 61.5 | 0 | 12.95 |
| ASearcher-Web-7B | 93.5 | 0 | 8.95 |

Table 12: Performance comparison of models on solution tasks showing Max Score, Min Score, and Macro Average.

| Model | Explicit Topic | Implicit Topic | Macro Avg. |
|---------------------------|----------------|----------------|------------|
| deepseek-V3.1 | 26.67 | 25.93 | 26.32 |
| OpenAI 4o-mini | 18.52 | 33.33 | 25.49 |
| kimi-k2 | 20 | 29.17 | 24.07 |
| Grok-4 | 20 | 20 | 20 |
| Gemini-2.5-flash-thinking | 13.33 | 13.33 | 13.33 |
| Qwen3-VL | 16.67 | 10 | 13.33 |
| Gemini-2.5-pro | 3.33 | 11.11 | 7.02 |
| MiroThinker-v1.0-30B | 3.7 | 4.17 | 3.92 |
| MMsearch-r1-7b | 3.7 | 3.7 | 3.7 |

Table 13: Performance comparison of models on Explicit Topic and Implicit Topic tasks.

| Task | Gemini-2.5 | GPT-5.1 | MM-r1 | Tulu-8b | Miro-30B |
|------------------|-------------|--------------|--------------|--------------|--------------|
| Reasoning | | | | | |
| Algorithm | 5 | 0 | 10 | 6 | 5 |
| Fig-Tab-Cha Com | 0 | 0 | 0 | 10 | 0 |
| Fig-Tab-Cha Rea | 10 | 0 | 5 | 5 | 5 |
| Formula | 10 | 0 | 20 | 0 | 5 |
| Paper | 10 | 0 | 5 | 0 | 0 |
| Induction | | | | | |
| Explicit | 13.3 | 0 | 3.7 | 0 | 3.7 |
| Implicit | 13.3 | 0 | 3.7 | 0 | 4.2 |
| Summary | | | | | |
| Trend | 41.0 | 50.6 | 44.2 | 51.8 | 29.8 |
| Dev. | 44.0 | 50.6 | 51.6 | 49.8 | 31.8 |
| Comparison | 52.2 | 55.2 | 49.8 | 49.8 | 28.8 |
| Solution | 29.7 | 12.95 | 17.21 | 38.71 | 32.02 |

Table 14: Performance comparison (Transposed View). Note: Model names are abbreviated for brevity.

Prompt used for Induction QA Construction

Task:

You are an expert researcher and data generator. Given a set of scientific papers, your goal is to generate search-style queries whose correct answer is exactly the provided set of papers as a whole, rather than any single paper..

Input Information:

- Paper Titles: {titles}
- Common Entities: {common_entities}

Generate two distinct queries where the answer is the provided list of papers.

1. Explicit Theme Query:

- Explicitly ask for papers or prior work.
- Clearly reflect the shared research theme implied by the paper titles and common entities.
 - Must incorporate the core concepts represented by the common entities, but avoid copying exact technical terms or phrases from titles or entities.
 - Use generalized, abstract, or paraphrased expressions (fuzzy matching) that focus on conceptual meaning rather than surface keywords.
 - The intent should be unambiguous, e.g., “find papers”, “look for studies”, “survey prior work”.

2. Implicit Theme Query:

- Do NOT explicitly ask for papers or literature.
- Embed the specific problem within a practical, real-world scenario (e.g., "Find works that can help me process long video").
 - The problem description should naturally require the methods, ideas, or solutions addressed collectively by the given papers.
 - Integrate the core theme implicitly, using high-level, generalized language instead of technical terminology.
 - The query should resemble how a practitioner or researcher would describe a real-world challenge without naming the underlying methods directly.

Output Requirement:

You must output a single valid JSON object containing exactly two fields: "explicit_query" and "implicit_query". Do not include any markdown formatting or explanation outside the JSON.

Example Output Format:

```
{{  
  "explicit_query": "Help me find research works that explore scalable approaches for improving reasoning and understanding across complex, multi-source data..",  
  "implicit_query": "I am building a system that needs to make sense of information spread across many documents and formats, and I want ideas that could help it reason reliably under such conditions..."  
}}
```

Figure 6: The prompts used for induction task QAs construction.

Prompt used for Solution QA Construction

Task:

You are an expert scientific information extractor. Your task is to faithfully and precisely extract structured knowledge from a given document that describes a complex technical task and its corresponding solution.

The input document already contains a complete solution. You must identify and extract the task definition, analytical reasoning, and technical solution exactly as presented in the document, without adding, inferring, or generalizing beyond the source text.

Extraction Requirements:

- All extracted content must be strictly grounded in the document; do not introduce external knowledge or assumptions.
- Preserve professional terminology, technical expressions, and specific values exactly as they appear.
- Do not paraphrase in a way that changes technical meaning.
- If information is not explicitly stated in the document, it must not appear in the output.
- The output must strictly follow the predefined JSON schema below. Missing or additional fields are not allowed.

Schema Definition and Field Constraints:

You must output a single valid JSON object with the following structure:

```
{
  "title": "", // The exact title of the document.
  "requirement": "Under these conditions of ..., complete the task of ...",
  // A detailed and comprehensive description of the task defined in the document.
  // It must include all explicit conditions, constraints, objectives, and assumptions
  // stated in the text.
  "solution": "This paper proposes the solution of ..., specifically, first ..., secondly ...",
  // A detailed, structured description of the solution proposed in the document.
  // The description should reflect the logical order presented in the text and explicitly
  // mention key methods, mechanisms, or stages.
  "analytical knowledge": [
    // Extract all analyses of task conditions and their implications from the document.
    {
      "idx": "analysis0",
      "condition": "Condition is ...", // A concrete condition, premise, or scenario explicitly
      // described in the document.
      "challenge": "Under this condition, conducting ... will face challenges such as ...,
      // which may lead to ..." // The challenges and potential negative consequences
      // explicitly analyzed in the document under this condition.
    }
  ]
}
```

Figure 7: The prompts used for solution task QAs construction.

Prompt used for Summary trend QA Construction

Task:

You are an expert scientific analyst. You are given {num_papers} scientific papers that share a set of common entities and collectively address a related research theme. Your task is to synthesize the information across these papers and extract the methodological development trends reflected in their content.

Input Information:

- Full text of {num_papers} scientific papers
- Common Entities: {common_entities}
- Query: {query}

Your task is to integrate this heterogeneous information and **summarize the development trend of the methods** discussed in these papers.

Extraction and Analysis Requirements:

- The summarized trends must be strictly grounded in the provided papers and input information.
- Do NOT fabricate, speculate, or introduce any information not explicitly supported by the documents.
- Trends must emerge from cross-paper comparison or temporal progression, not from a single paper in isolation.
- Preserve professional terminology and technical expressions as used in the papers.
- If temporal characteristics (e.g., chronological evolution, stage-wise progression) are present in the papers, they must be explicitly reflected in the trends.
- If no temporal signal exists, summarize trends in terms of conceptual or methodological shifts only.

Trend Characterization Guidelines:

The development trends may reflect, but are not limited to:

- Changes in methodological design or paradigm over time
- Increasing or decreasing scalability, efficiency, or generality
- Shifts in application domains or problem settings
- Evolution from heuristic or task-specific methods toward more unified or general approaches
- Increasing reliance on data scale, model capacity, or system-level integration

Output Requirements:

- Output must be a single valid JSON object.
- The output structure must strictly follow the provided example schema.
- Do not include markdown formatting, explanations, or any text outside the JSON.
- Do not add or remove fields beyond those defined in the example.

Common Entities:

{common_entities}

Query:

{query}

Output Format Example:

{example_json}

Figure 8: The prompts used for summary trend task QAs construction.

Summary Evaluation Prompt

Task:

You are an expert evaluator. Your task is to evaluate the quality of a model-generated response compared to a reference answer across five dimensions: Fluency, Relevance, Accuracy, Creativity, and Overall Quality.

Read both the reference answer and the model output, then assign a score from 1 to 5 for each dimension based on the detailed criteria below. Finally, calculate the average score across all five dimensions.

Scoring Dimensions and Criteria:

Fluency (1–5):

- 5: Grammatically flawless, natural, and easy to read.
- 4: Mostly fluent with minor grammatical or stylistic issues.
- 3: Understandable but occasionally awkward or unnatural phrasing.
- 2: Contains frequent grammatical or lexical errors that affect readability.
- 1: Poorly written, hard to understand, or ungrammatical.

Relevance (1–5):

- 5: Fully relevant; directly addresses all aspects of the reference answer or task.
- 4: Mostly relevant; minor omissions or small digressions.
- 3: Partially relevant; some key aspects are missing or off-topic.
- 2: Largely irrelevant; only a few parts relate to the task.
- 1: Completely irrelevant.

Accuracy (1–5):

- 5: Factually correct and logically consistent with the reference answer.
- 4: Minor factual or logical inaccuracies, but mostly correct.
- 3: Some inaccuracies or unsupported statements.
- 2: Multiple factual or reasoning errors.
- 1: Largely incorrect or misleading.

Creativity (1–5):

- 5: Demonstrates originality, insightful reasoning, or elegant rephrasing beyond the reference.
- 4: Shows moderate creativity while maintaining task alignment.
- 3: Standard or expected phrasing; minimal originality.
- 2: Repetitive or generic; little creative contribution.
- 1: Completely unoriginal or copied without adaptation.

Reference: {reference}

Model Output: {model_output}

Output Format (JSON):

Return your evaluation in the following JSON structure:

```
{example_output_json}
```

Notes: Each score should be an integer from 1 to 5.

average_score should be the mean of the five scores, rounded to one decimal place.

Provide concise, reasoned comments in the "comments" section explaining your scoring decisions and a brief "summary" of the overall evaluation.

Figure 9: The prompts used for summary task evaluation.

Solution Evaluation Prompt

Task:

```
{question}
<<Model-generated solution>>
{solution}
<<Judgement reference>>
## Analysis knowledge:
{analysis_knowledge}
## Technology knowledge:
{technology_knowledge}
## Golden explanation:
{golden_explanation}
## Golden solution:
{golden_solution}
<<Instruction>>
```

The above <<Task>> represents a complex real-world engineering requirement. The <<Model-generated solution>> is produced by a model and must be evaluated strictly against the <<Judgement reference>>, which is annotated by human experts. The Judgement reference consists of:

- (a) Analysis knowledge: a comprehensive enumeration and explanation of restrictive factors in the task.
- (b) Technology knowledge: a detailed description of the required or recommended technologies.
- (c) Golden explanation: expert reasoning on how technologies address the restrictive factors.
- (d) Golden solution: the expert-approved standard solution.

Your evaluation must fully rely on the Judgement reference and must not introduce external assumptions.

continued on next page

Figure 10: The prompts used for solution task evaluation.

| Failure Mode | Definition | Count | Ratio |
|-------------------|---|-------|-------|
| Extraction Error | Correct document retrieved, but unable to read precise numerical values from charts/tables. | 15 | 37.5% |
| Reasoning Error | Values extracted correctly, but multi-step calculation or logical inference fails. | 9 | 22.5% |
| Retrieval Failure | Fails to search (7.5%) or retrieves the wrong document (12.5%). | 8 | 20.0% |
| Hallucination | Outputs speculative guesses despite successful retrieval. | 6 | 15.0% |
| Context Limit | Key information truncated due to long multi-document concatenation. | 2 | 5.0% |

Table 15: Error attribution for Grok-4 based on 40 failure cases.

Solution Evaluation Prompt

<<Evaluation Requirements>>

1. You must evaluate the solution from **two dimensions**:

(1) Analysis Score; (2) Technology Score

Each score must be an integer between 0 and 100 (inclusive).

2. Scoring criteria:

(2.1) Analysis Score

Evaluate whether the <<Model-generated solution>> adequately considers the restrictive factors listed in the Analysis knowledge. You must:

- Enumerate each restrictive factor from the Analysis knowledge.
- Check whether the solution explicitly or implicitly addresses each factor.
- If addressed, determine whether the handling is correct and sufficiently specific.

Scoring rules:

- 0: No restrictive factors are considered.
- 11–30: Factors are mentioned but analysis is largely incorrect.
- 31–60: Factors are correctly considered but lack specificity.
- 61–90: Factors are correctly and specifically addressed, largely consistent with the Analysis knowledge.
- 100: Fully consistent with the Analysis knowledge and the Golden explanation.

(2.2) Technology Score

Evaluate whether the solution correctly employs the technologies listed in the Technology knowledge. You must:

- Enumerate each technology from the Technology knowledge.
- Check whether and where it is used in the solution.
- Assess correctness and specificity of its application.

Scoring rules:

- 0: No relevant technologies are used.
- 11–30: Technologies are used incorrectly.
- 31–60: Technologies are correctly used but lack specificity.
- 61–90: Technologies are used correctly and specifically, largely consistent with the Technology knowledge.
- 100: Fully consistent with the Technology knowledge and the Golden solution.

3. If Analysis knowledge, Technology knowledge, or Golden explanation is missing, base your evaluation primarily on similarity between the model-generated solution and the Golden solution in terms of analytical depth and technical correctness.

4. Length of the solution must not influence the score. Only correctness, coverage, and specificity relative to the Judgement reference matter.

5. **Output format constraint (strict):**

You must output **only** a JSON object in the following format, with no additional text, explanation, or reasoning:

```
{"Analysis Score": int,  
"Technology Score": int}
```

Figure 11: The prompts used for solution task evaluation.

A Correct Case Study of Grok-4 on PaperScope Bench Test

System:

System Prompt

User Query:

Across the ICLR 2025 papers, the methods that improve online continuous-control RL rely on the same training backbone, while the cross-domain offline adaptation method uses a different backbone. What are these two backbone algorithms??

Assistant:

<think>

****Thinking about the user's request**** ****Understanding the query**** - The question focuses on two backbone algorithms from ICLR 2025 papers in RL. ...

</think>

<tool_call>

```
{\"name\": \"FileSearchTool\", \"arguments\": {\"query\": \"ICLR 2025 papers on online continuous-control reinforcement learning improvements via explainability, action repetition, and cross-domain offline adaptation methods\", \"top_k\": 5}}
```

</tool_call>

User:

<tool_response>PDF Markdown Content+Images </tool_response>

Assistant:

<think>

****Thinking about the user's request**** ****Analyzing tool response**** The tool response brings up a paper titled \"Imitating Language via Scalable Inverse Reinforcement Learning\"

</think>

<answer>

The two backbone algorithms are Soft Actor-Critic (SAC) for the online continuous-control RL methods and Implicit Q-Learning (IQL) for the cross-domain offline adaptation method.

</answer>

Prediction: SAC, IQL

True answer: SAC, IQL

Figure 12: a case study of Grok-4.

A Wrong Case Study of Grok-4 on PaperScope Bench Test System:

System Prompt

User Query:

Across the collection, focusing on Walker2d in each study's setting, how much does the proposed method beat its primary baseline on the reported metric, and which method achieves the largest gain?

Assistant:

<think>

****Thinking about the user's request**** ****Analyzing query**** - The question targets performance comparisons in studies using the Walker2d environment,...A good query could be \"Walker2d proposed method baseline performance metric\" to pull up relevant files.

</think>

<tool_call>

```
{\"name\": \"FileSearchTool\", \"arguments\": {\"query\": \"Walker2d proposed method baseline performance metric comparison\", \"top_k\": 10}}
```

</tool_call>

User:

<tool_response>PDF Markdown Content+Images </tool_response>

Assistant:

<think>

****Thinking about the user's request**** ****Reviewing tool response**** - The tool response brings up a paper on \"Bisimulation Metric for Model Predictive Control\"...but none provide specific numerical performance metrics for Walker2d comparisons. - To get the needed details on how much proposed methods beat baselines,...

</think>

<tool_call>

```
{\"name\": \"FileSearchTool\", \"arguments\": {\"query\": \"Walker2d evaluation results performance metrics proposed method vs baseline DM Control D4RL MuJoCo\", \"top_k\": 10}}
```

</tool_call>

continued on next page

Figure 13: a case study of Grok-4.

A Wrong Case Study of Grok-4 on PaperScope Bench Test

User:

`<tool_response>`PDF Markdown Content+Images `</tool_response>`

Assistant:

`<think>`

****Thinking about the user's request**** ****Analyzing tool response**** - This paper discusses model-based reinforcement learning for improving sample efficiency in complex environments.- Since the available tools don't include the custom FileSearchTool used previously, switching to web_search for RL papers on Walker2d benchmarks. ...****Extracting gains from snippets**** - In the NUNO paper, the method outperforms state-of-the-art baselines by up to 93% on low-quality random datasets across all D4RL environments, ...

`</think>`

`<answer>`

The document corpus contains several recent RL papers evaluating on MuJoCo-based benchmarks. Walker2d specifically refers to the Walker2d-v2 (or similar) environment from OpenAI Gym MuJoCo, commonly used in the D4RL offline RL benchmark (with dataset variants such as random-v2, medium-replay-v2, medium-v2, medium-expert-v2, expert-v2). ****Summary of gains on Walker2d**** | Paper | Primary Baseline(s) | Walker2d Gain (approx.) | Dataset Type with Largest Gain | |---|-----|-----|-----| | VACO | IQL, TD3+BC, CQL, etc. | SOTA (exact Δ not stated) | All D4RL walker2d variants | | NUNO | MOPO, TATU+MOPO | ****Up to 93% relative**** | Low-quality/random | | Behavioral Entropy | Rényi/Shannon/RND/SMM | Superior (80–100% win rate, Δ not quantified) | Custom-generated | ****Method with the largest gain****: ****NUNO**** (Neural Stochastic Differential Equations for Uncertainty-aware Offline RL) achieves the largest reported improvement — ****up to 93% relative over prior SOTA**** on low-quality Walker2d tasks (random-v2 and equivalent low-coverage datasets).

`</answer>`

Prediction: 93, NUNO

True answer: 723, SDAR

Figure 14: a case study of Grok-4.