ARCHE: A Novel Task to Evaluate LLMs on Latent Reasoning Chain Extraction

Pengze Li[1, 2], **Jiaqi Liu**[2, 3], **Junchi Yu**[4], **Lihao Liu**[2], **Mingyu Ding**[3], **Wanli Ouyang**[2, 5], **Shixiang Tang**[2, 5*], **Xi Chen**[1, 6*]

[1]Artificial Intelligence Innovation and Incubation Institute of Fudan University
[2]Shanghai Artificial Intelligence Laboratory
[3]UNC-Chapel Hill
[4]University of Oxford
[5]The Chinese University of Hong Kong
[6]Shanghai Academy of AI for Science

Abstract

Large language models (LLMs) are increasingly used in scientific domains. While they can produce reasoning-like content via methods such as chain-of-thought prompting, these outputs are typically unstructured and informal, obscuring whether models truly understand the fundamental reasoning paradigms that underpin scientific inference. To address this, we introduce a novel task named Latent Reasoning Chain Extraction (ARCHE), in which models must decompose complex reasoning arguments into combinations of standard reasoning paradigms in the form of a Reasoning Logic Tree (RLT). In RLT, all reasoning steps are explicitly categorized as one of three variants of Peirce's fundamental inference modes: deduction, induction, or abduction. To facilitate this task, we release ARCHE Bench, a new benchmark derived from 70 Nature Communications articles, including more than 1,900 references and 38,000 viewpoints. We propose two logic-aware evaluation metrics: Entity Coverage (EC) for content completeness and Reasoning Edge Accuracy (REA) for step-by-step logical validity. Evaluations on 10 leading LLMs on ARCHE Bench reveal that models exhibit a trade-off between REA and EC, and none are yet able to extract a complete and standard reasoning chain. These findings highlight a substantial gap between the abilities of current reasoning models and the rigor required for scientific argumentation.

Code — https://github.com/Linsonng/ARCHEBenchmark/

1 Introduction

"All valid reasoning is either deductive, inductive, or hypothetic; or else it combines two or more of these characters."

— Charles S. Peirce

LLMs are increasingly applied in scientific domains, from assisting literature reviews (Wang et al. 2024; Zhu et al. 2025) and hypothesis generation (Gottweis et al. 2025; Xiong et al. 2024) to aiding in experimental design (Huang et al. 2024; Li et al. 2025). Although these advances suggest the potential of LLMs to accelerate scientific discovery(Bai et al. 2025), it is poorly understood how well these models understand and emulate human reasoning. In particular, their ability to follow and generate structured paradigm-based reasoning is uncertain, raising concerns about the trustworthiness of LLM-driven scientific workflows.

Building on Charles S. Peirce's taxonomy, which holds that all valid reasoning is deductive, inductive, abductive, or some combination thereof (Peirce 1868), we argue that the ability to understand and appropriately apply these elementary paradigms of reasoning is essential for LLMs to perform trustworthy scientific reasoning. Human scientists often employ a mixed inferential strategy to navigate vast bodies of evidence and competing claims in order to generate novel insights.

Such discoveries are not isolated leaps, but chains of reasoning steps that traverse multiple paradigms. This process involves latent reasoning chains composed of implicit, unspoken steps that connect existing knowledge to new insights. However, existing benchmarks fail to evaluate whether LLMs can (i) recognize the three reasoning paradigms from complex argument, (ii) incorporate them into coherent reasoning chains, and (iii) ground each step in verifiable textual evidence (Yang et al. 2024b; Hu et al. 2025).

To address limitations (i) and (ii), we propose a novel task, **Latent Reasoning Chain Extraction (ARCHE)**, designed to recognize the underlying reasoning behind scientific claims. ARCHE leverages an LLM to extract fine-grained reasoning steps from a paper's *introduction* paragraph and classify each step as deductive, inductive, or abductive reasoning. These steps are then assembled into a structured **Reasoning Logic Tree (RLT)**, where nodes correspond to individual premise or conclusion sentences, and labeled edges link each set of premise nodes to the corresponding conclusion node, indicating the associated inference type. By (i) recognizing distinct reasoning paradigms and (ii) assembling them into an RLT, ARCHE delivers a faithful, structured representation of complex scientific arguments.

To (iii) ground the reasoning steps in real tasks, we introduce **ARCHE Bench**, a benchmark derived from 70 peer-reviewed Nature Communication articles. For each paper,

*Corresponding author

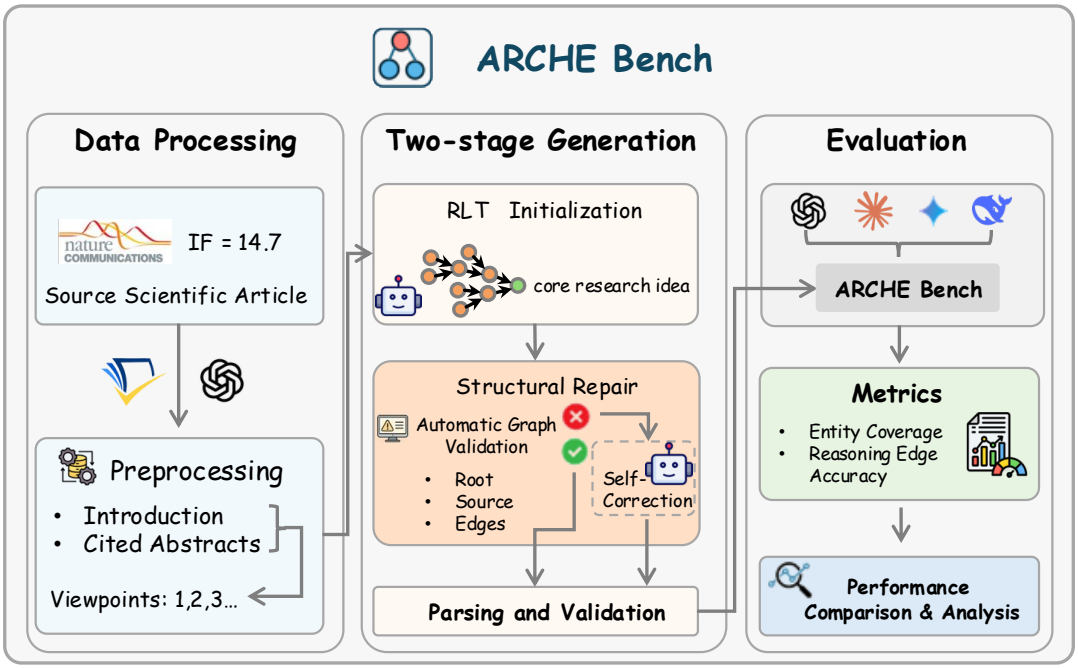

Figure 1: Overview of **ARCHE Bench**.The pipeline includes three stages: Data Processing, where scientific articles are preprocessed to extract *introductions*, cited abstracts, and viewpoints; RLT Generation, which constructs and repairs reasoning logic trees (RLTs) through automatic validation and self-correction; and Evaluation, where models are assessed using metrics like EC and REA for performance analysis.

we provide its *introduction* along with the relevant background viewpoints. We also propose two complementary metrics that capture different aspects of RLT: (a) EC, which measures proportion of key scientific entities of each paper that appear in the predicted RLT, reflecting how comprehensively the model captures the core contents; and (b) REA, which assesses the accuracy of each individual inference step in the reasoning chain, using an LLM-based judge to verify whether the conclusion logically follows from its premises given the labeled inference type. While EC ensures that no critical pieces of the core idea are omitted, REA directly evaluates logical validity at the step level.

We perform zero-shot evaluations on 10 state-of-the-art LLMs, revealing that even the best models struggle to exceed an accuracy of 50%, indicating their limited ability to recognize and properly formalize latent reasoning chains via standard paradigms. Furthermore, we observe a trade-off between EC and REA: models that achieve higher EC often do so at the cost of lower REA, and vice versa.

Our contributions are summarized as follows:

- We define a new task, Latent Reasoning Chain Extraction, to address the lack of evaluation on whether LLMs can model fundamental reasoning paradigms underlying human scientific reasoning.
- We design an automated pipeline for constructing ARCHE Bench, a dataset of scientific articles enriched with references and extracted viewpoints. We also introduce two evaluation metrics, EC and REA, to assess the completeness and correctness of generated RLTs.
- We benchmark 10 state-of-the-art LLMs and reveal that models often fail to extract and formalize reasoning chains in a structurally valid way. This highlights a fundamental gap between natural language fluency and paradigm-grounded scientific reasoning, distinguishing current LLM behavior from human expert inference.

2 Related Work

2.1 Reasoning in LLMs

Chain-of-Thought (CoT) prompting has become a foundational technique for eliciting multi-step reasoning in LLMs. Pioneered by (Wei et al. 2022) with few-shot examples and later shown to be effective in zero-shot settings (Kojima et al. 2022), CoT encourages models to generate intermediate reasoning steps. Subsequent work has enhanced its reliability through methods like self-consistency (Wang et al. 2023; Yu, He, and Ying 2024) and Tree-of-Thought (Yao et al. 2023). However, CoT produces an untyped narrative, lacking the formal logical grounding needed for verifiable reasoning. This limitation has spurred a move towards logic-aware systems. One direction is neuro-symbolic, where LLMs translate natural language into formal logic for an external solver, as seen in systems like LINC (Olausson et al. 2023). While formally robust, these methods can be brittle. Critically, most existing approaches study deduction, induction, and abduction in isolation, motivating our work on a unified framework.

2.2 LLMs for Scientific Discovery

In the scientific domain, LLMs are increasingly used to accelerate discovery, from domain-specific models that master literature like Galactica (Taylor et al. 2022) to systems that actively generate testable hypotheses or scientific inference(Luo 2025; Alkan et al. 2025; Wang et al. 2025b). Hybrid approaches combining LLMs with structured knowledge graphs have been particularly effective at generating novel and valid hypotheses (Tong et al. 2024). A parallel line of work focuses on evidence synthesis, where models must aggregate information to verify scientific claims (Wadden et al. 2020; Wan et al. 2025). To bring more structure to this process, the EntailmentBank benchmark requires models to construct deductive proof trees (Dalvi et al. 2021). However, these primarily focus on textual entailment or simple deduction. There remains a critical gap in evaluating an LLM's ability to perform complete scientific reasoning—integrating inductive generalization, deductive application, and abductive explanation within a single, coherent argument grounded in evidence. ARCHE Benchaddresses this gap.

2.3 Evaluation of Scientific Reasoning

The evaluation of LLM reasoning is a major challenge. Broad benchmarks like MMLU (Hendrycks et al. 2021) and specialized ones for math or logic (Cobbe et al. 2021; Yu et al. 2019; Wang et al. 2025a) typically focus on final-answer accuracy, which can mask an unsound reasoning process. Even when reasoning types are considered, they are often evaluated in separate, targeted benchmarks (Clark, Tafjord, and Richardson 2020; Yang et al. 2024a). In the scientific context, this gap persists; existing benchmarks do not require models to explicitly identify, differentiate, and compose all three Peircean inference types within a single, coherent argument. Our work takes a complementary approach: rather than outsourcing logic, we challenge LLMs to produce reasoning chains with explicit, typed inference steps. By requiring models to construct a graph of deductive, inductive, and abductive moves, ARCHE-bench provides a new lens for evaluation.

3 Methodology

In this section, we formalize the task of Latent Reasoning Chain Extraction (ARCHE) and describe the Reasoning-Logic Tree (RLT) designed to encode such reasoning. We then introduce ARCHE Bench, a dataset constructed specifically for this task.

3.1 Task Definition

The core objective of this task is to interpret complex paragraphs of scientific reasoning as a combination of individual premise viewpoints and standard reasoning paradigms.

The input of the task includes both the full *introduction* section of a scientific paper and *viewpoints* extracted from two sources: the *introduction* itself, and the *abstracts* of papers cited within that *introduction*. Following the definition by Feng et al. (Feng, Sun, and You 2025), a viewpoint refers to an idea, argument, or fact embedded within the research

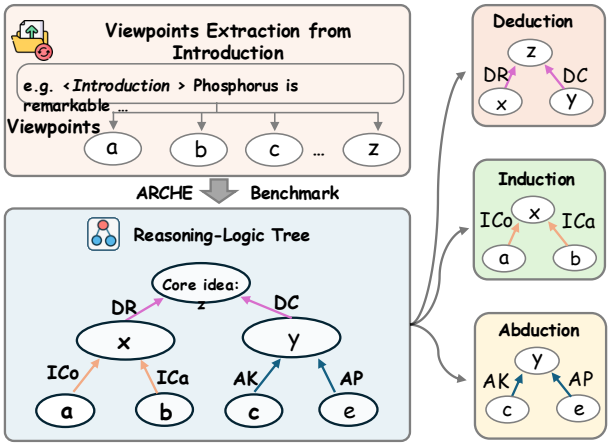

Figure 2: The Construction of Reasoning-Logic Tree. Viewpoints are first extracted from scientific text, then organized into a hierarchical reasoning structure. Each reasoning edge is annotated with an inference type: deduction, induction, or abduction, based on its logical pattern.

content. In our task, such viewpoints have already been extracted from the raw text using LLMs, prior to being input to any model.

The output RLT is a directed graph in DOT, a graph description language. A proper RLT should capture both the step-by-step inferential structure and the relevant viewpoints from the original paper or its references. Specifically, the model is expected to reconstruct how *asserted facts* (stated in the *introduction*), *established knowledge* (from cited abstracts) and *implicit domain knowledge* (unstated yet assumed by the author) are logically connected via elementary reasoning paradigms to derive the reasoning within the paper. This demonstrates whether the model has the capacity to extract the reasoning chain behind a scientific claim and to ground each inference step into a verifiable source, fulfilling the core objective of the ARCHE task.

3.2 Reasoning-Logic Tree (RLT)

RLT is a structured and logic-oriented graph representation designed to make implicit reasoning steps within scientific discourse explicit, intelligible, and machine-readable. In this section, we elaborate on its nodes, edges, and graph-level constraints, as well as the pipeline used for RLT generation.

Node Definition. Each node in the RLT contains a viewpoint and its source. The source of a viewpoint can be (i) a sentence or a viewpoint from the *Introduction* section, (ii) a viewpoint from a reference, or (iii) a sentence added by the LLM. In the third case, the added sentence often serves either as supplementary knowledge or implicit premises not explicitly stated by the author, or as an intermediate step required to bridge multi-hop reasoning. Further details of this coordinate system are provided in the appendix.

Edge Definition. The edges in an RLT indicate the reasoning paradigms underlying each inference. Each edge is directed and labeled, connecting a single premise node to a

conclusion node. We define six specific edge types, which serve as fine-grained instances of Peirce's three classical paradigms of reasoning: deduction, induction, and abduction. In practice, a reasoning step should involve multiple edges, connecting at least two premise nodes to a shared conclusion node. The six edge types are formally defined below.

- **Deduction-Rule (DR):** A premise stating a general principle, law, or established rule.
- **Deduction-Case (DC):** A premise that presents a specific instance or case that falls under the general rule.
- **Induction-Common (ICo):** A premise that abstracts a common pattern or regularity across multiple cases.
- **Induction-Case (ICa):** A premise providing a specific observation.
- **Abduction-Knowledge (AK):** A premise stating existing knowledge or known mechanisms.
- **Abduction-Phenomenon (AP):** A premise describing an observation requiring explanation.

Graph-Level Structural Constraints. To ensure that each RLT captures a coherent and focused reasoning path, we impose the following structural constraints. Each RLT is structured as a single-rooted directed acyclic graph (DAG). All nodes in the graph must converge at the central node, and disconnected or irrelevant premises are strictly prohibited. Furthermore, we enforce a standardized inference granularity by requiring that each logical step corresponds to a single reasoning paradigm, represented using one pair from the six predefined edge types. In cases of multi-hop reasoning, LLM is expected to introduce intermediate nodes that decompose the reasoning chain into multiple steps, each aligned with a specific paradigm. Edges may also cross document boundaries. For example, an edge may point from a viewpoint in a cited reference to one in the main paper, demonstrating how prior work provides premises for new research.

RLT Generation Pipeline. We design a fully automated two-stage pipeline to guide the LLM in generating a valid RLT from the source paper:

1. **Stage 1: Primary Extraction.** The LLM is first guided by a manually crafted prompt to extract the latent reasoning chain and produce an initial graph in the DOT language. The main ideas from Section 3.1 and Section 3.2 are included in the prompt. The full prompts are provided in the appendix.
2. **Stage 2: Structural Repair.** A verifier script automatically inspects the LLM-generated RLT for structural defects, such as multiple roots, cycles, disconnected nodes, or invalid edge labels. If no defects are found, this stage is skipped. Otherwise, the same LLM is re-prompted with a targeted prompt that instructs it to correct the structural errors while preserving the original reasoning content.

Algorithm 1: Evaluation for EC and REA

Input: paper text P, generated RLT
Output: EC, REA

1: core_entities $\leftarrow$ ExtractCoreEntities(P)
2: reasoning_steps $\leftarrow$ ParseReasoningSteps(RLT)
3: **for** each step $s \in$ reasoning_steps **do**
4: **if** ValidFormat(s) **then**
5: votes $\leftarrow$ ThreeModelVoting(s)
6: validation_results$[s] \leftarrow$ MajorityVote(votes)
7: **else**
8: validation_results$[s] \leftarrow$ format_error
9: **end if**
10: **end for**
11: $\mathrm{REA} \leftarrow \frac{|\{\text{s:validation\_results[s]="correct"}\}|}{|\text{reasoning\_steps}|}$
12: correct_nodes $\leftarrow \{n : n \in s, \text{validation\_results}[s] = \text{"correct"}\}$
13: reasoning_entities $\leftarrow$ ExtractEntities(correct_nodes)
14: $\mathrm{EC} \leftarrow \frac{|\text{core\_entities} \cap \text{reasoning\_entities}|}{|\text{core\_entities}|}$
15: **return** $(\mathrm{EC}, \mathrm{REA})$

3.3 ARCHE Bench

Building on the ARCHE task and the RLT representation detailed above, we introduce **ARCHE Bench**, a benchmark designed to evaluate LLMs on real scientific texts. It consists of 70 peer-reviewed articles published in *Nature Communications* in 2025. For each article, the *Introduction* section is extracted, along with viewpoints obtained from the *introduction* itself and from the abstracts of its cited references, using the Semantic Scholar API and GPT-4o. To support the evaluation, we propose an automated evaluation framework (Algorithm 1) along with two complementary metrics.

Dataset Construction and Benchmark Statistics. For each article in ARCHE Bench, the *introduction* section is extracted as the primary input for Stage 1 of RLT generation. All sentences from the *introduction* and from the abstracts of cited references are parsed into viewpoints using GPT-4o. Each sentence and its corresponding viewpoints are indexed to support precise referencing during generation. The ARCHE bench comprises 70 peer-reviewed and open-access articles published in 2025, ensuring that the content is up-to-date and scientifically verified. The benchmark is balanced across two major scientific domains: Biological Sciences (35) and Physical Sciences (35). In total, the corpus contains 2,164 sentences from the articles' *introductions* and 1,891 cited references, from which over 38,000 distinct viewpoints have been extracted for reasoning analysis. On average, each article's *introduction* provides 30.9 sentences and 77.4 viewpoints, and is supported by 27 citations. More details are provided in Table 2.

Evaluation. The algorithm 1 outlines the evaluation procedure for computing the EC and REA metrics. In the following, we elaborate on the key steps and underlying logic for each metric.

Entity Coverage (EC). This metric assesses the extent to which the generated RLT captures the core scientific concepts of the input article. Lines 1, 9–11

Model	Physical Sciences		Biological Sciences		Overall	
	REA(↑)	EC(↑)	REA(↑)	EC(↑)	REA(↑)	EC(↑)
Claude-Opus-4 (Thinking)	25.5%	68.1%	22.9%	71.2%	24.2%	**69.7%**
Claude-Sonnet-4 (Thinking)	28.6%	47.2%	29.0%	58.9%	28.8%	53.1%
DeepSeek-R1	16.3%	31.7%	24.0%	25.6%	20.1%	28.7%
Doubao-Seed-1.6 (Thinking)	22.6%	55.6%	**46.2%**	54.3%	28.2%	55.3%
Gemini-2.5-Pro	**38.4%**	58.8%	40.5%	54.5%	39.5%	56.7%
Gemini-2.5-Pro (Thinking)	38.0%	49.4%	44.9%	59.2%	**41.4%**	54.1%
GPT-4o	12.5%	28.7%	19.1%	19.8%	15.8%	24.3%
Grok-3	36.0%	63.3%	30.2%	44.3%	33.1%	53.8%
Grok-4	24.3%	55.7%	19.9%	**68.2%**	22.2%	61.7%
o3	32.3%	**64.4%**	44.3%	50.2%	35.6%	60.5%

Table 1: Performance Comparison of LLMs on ARCHE.

of Algorithm 1 describe how EC is calculated. First, `ExtractCoreEntities(P)` uses the o3 model (OpenAI 2025) to extract the core scientific idea from the *introduction* of the article, identifying the main hypothesis or methodology. From this, scientific entities, concrete concepts, methods, and phenomena are extracted (typically 8-10). Coverage is calculated as the proportion of extracted entities that appear in the nodes involved in correct reasoning steps in the generated RLT. Entity matching is performed using case-insensitive string comparison.

Reasoning Edge Accuracy (REA). This metric measures the percentage of individual reasoning steps in the generated RLT that are logically valid. Lines 2–8 compute REA by validating each reasoning step in the generated RLT. Evaluation begins by identifying the root node and collecting all reasoning paths connected to it. Each step is validated for format and categorized into reasoning types. Steps with inconsistent or invalid combinations, such as pairing a deduction case with an abduction-knowledge edge, or using non-standard labels like 'deduction-knowledge', are automatically marked incorrect (line 7). Valid steps are evaluated using a three-model voting system (`ThreeModelVoting`, line 5), where *o3* (OpenAI 2025), *Claude-Sonnet-4-thinking* (claude 2025), and *Gemini 2.5 Pro* (Team 2025) independently judge whether the conclusion follows logically from the premises. A majority vote determines the final label for each step (line 6). REA is then computed as the proportion of reasoning steps labeled as correct (line 8). We evaluated the voting system across three reasoning types using human annotations, finding that the joint model consistently outperforms individual models in alignment with human judgments, achieving an accuracy exceeding 88%.

4 Experiments

4.1 Models

We evaluated 10 leading LLMs that span six model families. These include *Claude-Opus-4-thinking* (claude 2025), *Claude-Sonnet-4-thinking* (claude 2025), *GPT-4o* (OpenAI 2024), *o3* (OpenAI 2025), *DeepSeek-R1* (DeepSeek-AI and Guo 2025), *Gemini-2.5-Pro* (Team 2025), *Gemini-2.5-Pro-thinking* (Team 2025), *Grok-3* (xAI 2025a), *Grok-4* (xAI 2025b), and *Doubao-Seed-1.6-thinking* (ByteDance 2025).

Overall	
Total Articles	70
Total Sentences	2,164
Total Viewpoints	5,418
Total Citations	1,891
Total Referenced Viewpoints	33,321
Total Viewpoints (Combined)	**38,739**
Average per Article	
Sentences	30.9
Viewpoints	77.4
Citations	27.0
Viewpoints per Sentence	2.5
Publication Year	
2025	70

Table 2: ARCHE Bench Overall and Average Statistics.

All models are evaluated in the ARCHE task (Section 3.1) under same conditions, using the same prompts for consistency. Temperature is fixed at 0.1 when configurable. We report performance using the EC and REA metrics described in Section 3.3.

4.2 Performance Comparison and Analysis

The overall results show that the models demonstrate moderate performance across both evaluation dimensions.

The average EC is 51.4%, with a wide spread from 0% to 100%. The median is 66.7%, indicating that the models are able to identify most of the core scientific entities, despite high variance. REA is relatively low, with an average of 28.3% and a median of 25%. Additional results are provided in Table 1.

As shown in Figure 3, no model performs ideally in both EC and REA. A perfect RLT should demonstrate 100% accuracy in tracing reasoning steps from the given viewpoints to the core idea, while simultaneously covering all key entities, as illustrated by the green region in the figure.

In terms of overall performance, models with advanced reasoning capabilities, such as *Gemini-2.5-Pro-thinking*, *o3*, *Grok-3*, and *Claude-Opus-4-thinking*, perform significantly

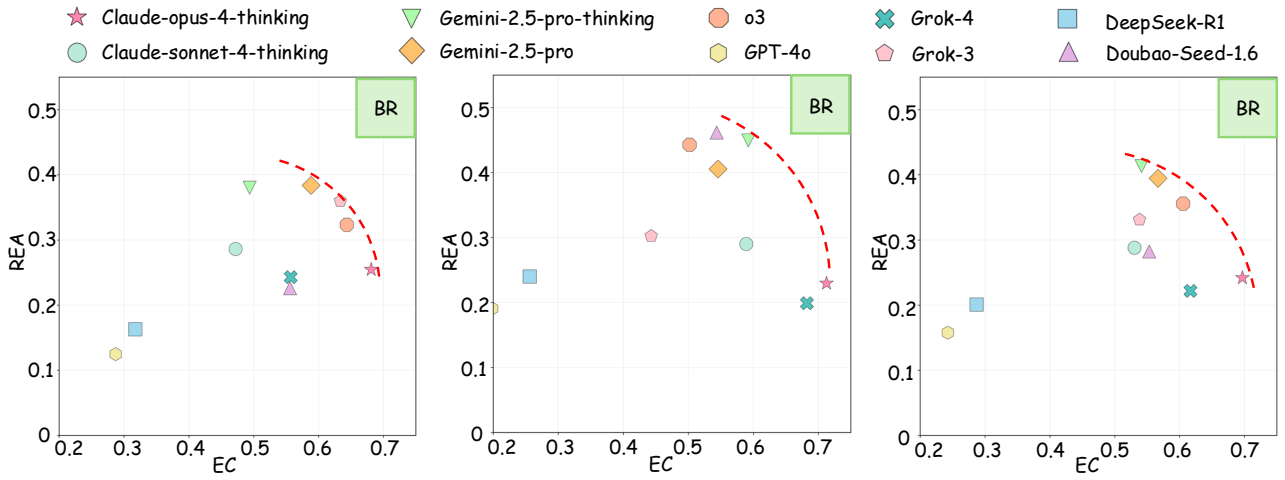

Figure 3: Comparative performance of LLM models in terms of EC and REA across three domains: (Left) Physical Sciences, (Middle) Biological Sciences, and (Right) Overall. Each point represents a model's performance. The green region (BR) indicates a preferable area with both higher coverage and accuracy. The red dashed curve denotes the trade-off frontier.

better than comparatively less optimized models like *gpt-4o*. This suggests that existing techniques, such as chain-of-thought prompting and reasoning-specific optimization strategies, indeed play a meaningful role in enhancing model reasoning. These observations also indirectly validate the utility of the ARCHE benchmark. For example, *o3* significantly outperforms *GPT-4o* on both EC and REA, while *Claude-Opus-4* achieves much higher EC scores than *Claude-4*. On the other hand, models with similar architectures and training scales exhibit comparable behavior. For example, *Gemini-2.5-Pro* and its thinking variant show only minor differences, with the latter producing a slight improvement in REA but no substantial overall gain.

However, the highest-performing models interestingly appear to align along a smooth curve (highlighted by the red dotted line), indicating that despite differences in model architecture and training objectives, there exists an apparent boundary in the joint space of accuracy and coverage. From this perspective, no single model has yet demonstrated a substantially superior ability to master and formalize reasoning chains with the fundamental paradigms.

4.3 Reasoning Performance by Type

Table 3 presents the performance of the model in three types of reasoning: abductive, deductive, and inductive. *Grok-3* emerges as the model that performs the best overall, achieving the highest scores in all categories. *GPT-4o* and *Gemini-2.5-Pro* also show strong results, particularly in deductive reasoning. *Claude-Sonnet-4* performs well on abductive tasks, with the accuracy reaching 63.7%.

In contrast, models such as *DeepSeek-R1* and *o3* show relatively weak performance in deductive reasoning, both scoring around 40%. Although *Grok-4* belongs to the same family as *Grok-3*, its performance is notably lower in both deductive and inductive tasks. This discrepancy is likely due to *Grok-4* generating more outputs without providing valid reasoning, which reduces step-level accuracy.

Model	Abductive	Deductive	Inductive
Claude-Opus-4	58.9%	42.4%	57.1%
Claude-Sonnet-4	62.0%	50.0%	63.7%
DeepSeek-R1	48.8%	40.6%	59.0%
Doubao-Seed-1.6	54.1%	46.9%	48.0%
GPT-4o	56.9%	63.4%	59.3%
Gemini-2.5-Pro	60.3%	59.5%	56.7%
Gemini-2.5-Pro-thinking	72.5%	56.9%	55.5%
Grok 3	87.1%	74.0%	77.9%
Grok 4	58.3%	36.6%	40.0%
o3	57.4%	40.0%	42.2%

Table 3: Reasoning Accuracy on 3 Types of Tasks.

Note that Table 3 only includes reasoning steps with valid output formatting. As a result, accuracy values are substantially higher than those reported in Table 1, which also accounts for formatting errors. Despite prompt refinement and multiple rephrasings, formatting issues remain prevalent. This suggests that the current setup already approaches the expressive limits of these models. The root cause lies in their inability to grasp the structural constraints of the reasoning paradigms and to express them in syntactically valid forms. This also explains why *GPT-4o* does not appear significantly weaker than other models in Table 3. A large portion of its output contains structural violations, such as mixing incompatible reasoning types, which are filtered out during the format validation. This further demonstrates the robustness of our benchmark, which reliably distinguishes models by their ability to generate structurally valid and correct reasoning, rather than surface-level fluency.

4.4 Analysis of Step Efficiency

Table 4 compares the extraction behavior of reasoning chains of different language models, using Average Total Steps (ATS) and Average Effective Steps (AES) as evalu-

Model	ATS	AES
Gemini-2.5-Pro	12.4	5.8
Gemini-2.5-Pro-thinking	13.2	5.3
o3	11.7	4.9
Grok-4	20.1	4.9
Grok-3	11.0	4.0
Claude-Opus-4	11.0	3.3
Doubao-Seed-1.6	8.6	2.5
Claude-Sonnet-4	8.1	2.2
DeepSeek-R1	8.9	1.9
GPT-4o	9.2	1.2

Table 4: Performance Comparison of Language Models by Step Metrics. ATS: Average Total Steps; AES: Average Effective Steps.

ation metrics. ATS reflects the average number of reasoning steps generated per RLT, while AES measures the subset of steps that are logically valid and relevant to the final answer.

The results show that Grok-4 has the highest ATS (20.1), indicating a tendency to produce longer reasoning chains. However, its AES (4.9) does not increase proportionally, resulting in a relatively low proportion of effective steps. In contrast, o3 achieves a similar AES (4.9) with a much shorter ATS (11.7). A similar pattern is observed in Gemini-2.5-Pro-thinking, which generates slightly more steps (ATS 13.2) than the base Gemini-2.5-Pro (ATS 12.4), but with a marginally lower AES (5.3 vs. 5.8). Models such as Claude-Sonnet-4 and DeepSeek-R1 exhibit lower ATS values (around 8–9), with correspondingly low AES scores. In particular, GPT-4o produces the lowest AES (1.2) among all models, suggesting that it typically contributes only a single effective reasoning step per instance.

These findings indicate that longer reasoning chains do not necessarily correlate with higher reasoning quality. Some models tend to produce verbose or redundant outputs without improving step utility. Furthermore, even the best-performing model extracts fewer than six valid inferences on average from *introductions* containing more than 30 sentences, despite also having access to supporting viewpoints from references. This highlights a substantial gap between the performance of the current model and the goal of robust paradigm-aligned scientific reasoning.

5 Discussion

5.1 Key Findings and Insights

Our experiments reveal a fundamental limitation in the ability of current LLMs to explain scientific reasoning. Although many leading large reasoning models are capable of generating fluent, reasoning-like natural language outputs, they still lack a genuine grasp of reasoning itself. Even when presented with peer-reviewed scientific texts that contain verified and logically coherent points of view, these models often fail to identify the underlying logic of reasoning behind scientific claims and evidence.

This deficiency poses a major risk in the use of LLMs for scientific discovery. Scientific discovery requires a transparent and grounded thought process for rigorous verification and supervision. To this end, we advocate for the explicit incorporation of reasoning-paradigm-aligned data during model pre-training and instruction. Alternatively, reasoning supervision objectives could be augmented with reward signals grounded in formal paradigms to promote structurally valid and interpretable inference.

This motivation underlies the name of our task, **ARCHE**. The term *arch*, originating from ancient Greek philosophy and later formalized by Aristotle, denotes 'principle', 'source' or 'cause'. It reflects our belief that an effective reasoning benchmark must go beyond surface-level correctness and instead guide the development of models capable of verifiable, paradigm-based reasoning in complex scientific domains. We hope that the ARCHE Bench will serve as a foundation for future work toward models that not only speak in the language of reasoning but also reason in its true form.

5.2 Limitations and Future Work

Despite the insights obtained, our study has several limitations, which we aim to address in future work:

1. **Limited corpus scale.** The current benchmark includes only 70 research articles. Each ARCHE evaluation requires the model to process a full *introduction* along with all cited abstracts, resulting in high token usage, approximately $4 per paper across all 10 models. Although this constrains the size of the dataset, previous work (Press et al. 2024) suggests that small, carefully curated corpora can still meaningfully reveal model weaknesses. To improve both scale and diversity, we plan to expand the benchmark to include multidisciplinary content such as chemistry and artificial intelligence to assess cross-domain generalization.
2. **Partial research context.** The current benchmark focuses solely on the *Introduction* section, as it typically presents the central hypothesis or contribution. However, excluding *Methods* and *Results* may underestimate an LLM's ability to reason across the full scientific workflow. Reasoning based on experimental findings and iterative hypothesis formation is also essential. We will extend ARCHE to include these sections, allowing end-to-end evaluation of hypothesis generation, evidence collection, and conclusion formation.

6 Conclusion

We present a novel task, ARCHE, along with an automated benchmark designed to evaluate LLMs on scientific reasoning. Through evaluation on ARCHE Bench, we demonstrate that state-of-the-art models, including those optimized for reasoning, still struggle to reliably identify and organize reasoning chains in scientific texts. Our multi-perspective analysis reveals that fluent natural language reasoning does not imply an internalized understanding of reasoning paradigms. These findings highlight the need for paradigm-guided reasoning supervision, whether through data, reward design, or training framework. We hope that ARCHE will serve as a foundation for future research toward models that truly learn and apply reasoning.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory. This work was done during Pengze Li's internship at Shanghai Artificial Intelligence Laboratory. The computations in this research were performed using the CFFF platform of Fudan University.

References

Alkan, A. K.; Sourav, S.; Jablonska, M.; Astarita, S.; Chakrabarty, R.; Garuda, N.; Khetarpal, P.; Pióro, M.; Tanoglidis, D.; Iyer, K. G.; Polimera, M. S.; Smith, M. J.; Ghosal, T.; Huertas-Company, M.; Kruk, S.; Schawinski, K.; and Ciucă, I. 2025. A Survey on Hypothesis Generation for Scientific Discovery in the Era of Large Language Models. *arXiv preprint arXiv:2504.05496*.

Bai, L.; Cai, Z.; Cao, Y.; Cao, M.; Cao, W.; Chen, C.; Chen, H.; Chen, K.; Chen, P.; Chen, Y.; et al. 2025. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

ByteDance. 2025. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/seed1_6.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. arXiv:2002.05867.

claude. 2025. Introducing Claude 4. https://www.anthropic.com/news/claude-4.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Dalvi, B.; Jansen, P.; Tafjord, O.; Xie, Z.; Smith, H.; Pipatanangkura, L.; and Clark, P. 2021. Explaining Answers with Entailment Trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7358–7370.

DeepSeek-AI; and Guo, D. e. a. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Feng, T.; Sun, Y.; and You, J. 2025. Grapheval: A lightweight graph-based llm framework for idea evaluation. *arXiv preprint arXiv:2503.12600*.

Gottweis, J.; Weng, W.-H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hu, M.; Ma, C.; Li, W.; Xu, W.; Wu, J.; Hu, J.; Li, T.; Zhuang, G.; Liu, J.; Lu, Y.; et al. 2025. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Huang, K.; Qu, Y.; Cousins, H.; Johnson, W. A.; Yin, D.; Shah, M.; Zhou, D.; Altman, R.; Wang, M.; and Cong, L. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 22199–22213.

Li, J.; Chen, Y.; Liu, C.; Cai, Q.; Liu, T.; Han, B.; Zhang, K.; and Xiong, H. 2025. Can Large Language Models Help Experimental Design for Causal Discovery? *arXiv preprint arXiv:2503.01139*.

Luo, X. e. a. 2025. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 9: 305–315.

Olausson, T. X.; Gu, A.; Lipkin, B.; Zhang, C. E.; Solar-Lezama, A.; Tenenbaum, J. B.; and Levy, R. P. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/.

Peirce, C. S. 1868. Some consequences of four incapacities. *The Journal of Speculative Philosophy*, 2(3): 140–157.

Press, O.; Hochlehnert, A.; Prabhu, A.; Udandarao, V.; Press, O.; and Bethge, M. 2024. CiteME: Can Language Models Accurately Cite Scientific Claims? *Advances in Neural Information Processing Systems*, 37: 7847–7877.

Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A Large Language Model for Science. *arXiv preprint arXiv:2211.09085*.

Team, G. 2025. Gemini 2.5: Our most intelligent AI model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/#gemini-2-5-thinking.

Tong, S.; Mao, K.; Huang, Z.; Zhao, Y.; and Peng, K. 2024. Automating psychological hypothesis generation with AI: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11: 896.

Wadden, D.; Lin, S.; Lo, K.; Wang, L. L.; van Zuylen, M.; Cohan, A.; and Hajishirzi, H. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550.

Wan, H.; Yang, C.; Yu, J.; Tu, M.; Lu, J.; Yu, D.; Cao, J.; Gao, B.; Xie, J.; Wang, A.; et al. 2025. DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks. *arXiv preprint arXiv:2509.01396*.

Wang, L.; Su, E.; Liu, J.; Li, P.; Xia, P.; Xiao, J.; Zhang, W.; Dai, X.; Chen, X.; Meng, Y.; et al. 2025a. PhysUniBench: An Undergraduate-Level Physics Reasoning Benchmark for Multimodal Models. *arXiv preprint arXiv:2506.17667*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*.

Wang, Y.; Guo, Q.; Yao, W.; Zhang, H.; Zhang, X.; Wu, Z.; Zhang, M.; Dai, X.; Zhang, M.; Wen, Q.; et al. 2024. Autosurvey: Large language models can automatically write surveys, 2024. *arXiv preprint arXiv:2406.10252*.

Wang, Y.; Tang, C.; Deng, H.; Xiao, J.; Liu, J.; Wu, J.; Yao, J.; Li, P.; Su, E.; Wang, L.; et al. 2025b. SciReasoner: Laying the Scientific Reasoning Ground Across Disciplines. *arXiv preprint arXiv:2509.21320*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 24824–24837.

xAI. 2025a. Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3.

xAI. 2025b. Grok 4. https://x.ai/news/grok-4.

Xiong, G.; Xie, E.; Shariatmadari, A. H.; Guo, S.; Bekiranov, S.; and Zhang, A. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*.

Yang, Z.; Dong, L.; Du, X.; Cheng, H.; Cambria, E.; Liu, X.; Gao, J.; and Wei, F. 2024a. Language Models as Inductive Reasoners. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 209–225. St. Julian's, Malta: Association for Computational Linguistics.

Yang, Z.; Du, X.; Mao, R.; Ni, J.; and Cambria, E. 2024b. Logical Reasoning over Natural Language as Knowledge Representation: A Survey. arXiv:2303.12023.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:arXiv:2305.10601.

Yu, J.; He, R.; and Ying, Z. 2024. THOUGHT PROPAGATION: AN ANALOGICAL APPROACH TO COMPLEX REASONING WITH LARGE LANGUAGE MODELS. In *The Twelfth International Conference on Learning Representations*.

Yu, W.; Jiang, Z.; Dong, Y.; and Feng, J. 2019. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *International Conference on Learning Representations (ICLR)*.

Zhu, M.; Weng, Y.; Yang, L.; and Zhang, Y. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.