SciCompanion: Graph-Grounded Reasoning for Structured Evaluation of Scientific Arguments

Joshua Flashner[1], **Adithya Kulkarni**[2], **Dawei Zhou**[1]

[1] Department of Computer Science, Virginia Tech
[2] Department of Computer Science, Ball State University
jflashner@vt.edu, adithya.kulkarni@bsu.edu, zhoud@vt.edu

Abstract

The exponential growth of scientific publications has overwhelmed reviewers and researchers, with top conferences receiving thousands of submissions annually. Reviewers must assess feasibility, novelty, and impact under tight deadlines, often lacking tools to identify relevant prior work. Early-career researchers face similar challenges, with limited support to navigate fast-evolving fields. Existing LLM-based systems struggle with static retrieval, surface-level features, and lack multi-hop reasoning, leading to shallow or hallucinated assessments. Scientific evaluation requires a deep, relational understanding, which current retrieval-augmented generation (RAG) methods fail to achieve. We introduce SciCompanion, a graph-grounded reasoning framework for structured scientific evaluation. Given a paper or abstract-like input, SciCompanion builds a dynamic knowledge graph from recent publications, domain-specific databases, and curated metadata. It employs multi-hop reasoning to iteratively construct contextual graphs and generate structured critiques, enabling deeper exploration of scientific literature. Unlike sentiment-biased LLM evaluations, SciCompanion directly optimizes retrieval and graph refinement using Group Relative Policy Optimization (GRPO), producing reviews aligned with expert judgments. Experiments on ICLR and ACL datasets show that SciCompanion reduces evaluation error by over 30% compared to prompting-only baselines and allows smaller models to outperform larger ones. Evaluations across three datasets, using metrics for retrieval accuracy, semantic overlap, and multi-hop sensitivity, along with a case study, demonstrate SciCompanion's robustness and versatility.

1 Introduction

The exponential rise in scientific publications has immensely strained the peer review ecosystem. Conferences in artificial intelligence and machine learning, such as NeurIPS, ICML, and ICLR, have seen a significant increase in paper submissions, with NeurIPS 2025 receiving over 10,000 submissions (Xu et al., 2024). Similarly, ACL conferences have experienced consistent year-over-year growth, with ACL 2023 reporting 4,864 submissions, a marked increase from previous cycles (Bharti et al., 2023). This surge creates unsustainable reviewer workloads due to high volume, tight deadlines, and unfamiliarity with subdomains (Mehmani and Ghildiyal, 2024). The "publish or perish" culture (Guraya et al., 2016) exacerbates this, encouraging quantity over rigor and leading to reviewer fatigue.

Early-career researchers and junior reviewers also struggle with the rapidly growing, fragmented literature (Johnson and Weivoda, 2021; Bandichhor et al., 2023). As prior work exceeds individual cognitive capacity, assessing novelty, identifying related work, and evaluating methodology becomes time-consuming and error-prone. This information overload compromises peer review quality and scientific judgment, highlighting the urgent need for intelligent, scalable, and trustworthy tools for transparently synthesizing, contextualizing, and evaluating contributions (Picano, 2025).

Large language models (LLMs) offer scalable language understanding but falter in the face of evolving, frontier scientific knowledge (Ye et al., 2024; Zeng et al.). In an attempt to resolve this, Retrieval-Augmented Generation (RAG) approaches (Lewis et al., 2020; Liu, 2025) incorporate external documents, but are typically *static*, *non-adaptive*, and *unstructured* (Barnett et al., 2024; Han et al., 2025a). Graph-based methods like GraphRAG (Han et al., 2025b,a) offer structured retrieval, yet they typically focus on passive information linkage rather than *critique-driven* synthesis or *task-conditioned* reasoning. LLM baselines lack alignment with expert review dimensions, of-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24223–24244
November 4-9, 2025

Property	GPT-4	PeerRead	OpenReviewer	PaperSEA	SciCompanion
Comprehensive Validation	✓				✓
Multi-hop Retrieval					✓
Task Specific Optimization		✓	✓	✓	✓
Cross-domain Adaptability	✓			✓	✓

Table 1: Comparison of SciCompanion with baseline and specialized peer review systems. SciCompanion is the only framework that integrates dynamic, multi-hop retrieval with critique-aligned optimization.

ten hallucinate unsupported claims (Ji et al., 2023), and fail to support multi-hop reasoning. For instance, standard LLMs may retrieve superficially relevant papers but fail at the multi-hop reasoning needed to uncover subtle connections that determine true novelty, as they are not inherently designed for deep, iterative exploration. Recent specialized systems have sought to address these gaps: OpenReviewer (Idahl and Ahmadi, 2025) uses large-scale fine-tuning on expert reviews to generate more critical and realistic feedback that matches human judgment distributions, while PaperSEA (Yu et al., 2024) introduces a multi-stage framework to standardize inconsistent review data before fine-tuning and apply a self-correction mechanism. However, these approaches are primarily grounded in supervised learning from existing text and do not incorporate dynamic, graph-based reasoning or reinforcement learning to guide the critique generation process.

Effective peer review demands systems for deep, context-sensitive evaluation that are: context-aware (interpreting domain nuance), critique-aligned (structured around feasibility, novelty, impact), and explainable (producing interpretable, trustworthy reasoning) (Bharti et al., 2023; Kumbhar et al., 2025; Xiong et al., 2024). Prior symbolic and graph-based tools (Ji et al., 2021; Dessì et al., 2021; Oelen et al., 2020) offer structured exploration but are disconnected from modern LLMs' adaptive reasoning and lack reinforcement learning scaffolds for alignment with scientific critique (Lu et al., 2024), leaving a gap for structurally grounded, flexible systems.

To address the limitations of static retrieval and shallow critique in scientific evaluation, we introduce SciCompanion, a unified framework that integrates dynamic graph reasoning, reinforcement learning, and LLM-driven critique generation (see Table 1). Departing from conventional RAG pipelines and static knowledge graph systems, SciCompanion builds evolving, multi-hop graphs grounded in scientific text, continuously refined through reinforcement signals via Group Relative Policy Optimization (GRPO) (DeepSeek-AI and et al., 2024; Schulman et al., 2017; Silver et al., 2018). This enables the system to adaptively retrieve, link, and assess evidence based on task-specific prompts. Each reasoning trajectory is explicitly aligned with structured review criteria, feasibility, novelty, and impact, drawing from advances in multi-agent prompting (Kumbhar et al., 2025), graph-centric LLM interfaces (Li et al., 2024), and scientific QA pipelines (Lu et al., 2022). Flexible, SciCompanion operates on full papers or abstract-like descriptions, supporting reviewers with heavy loads and researchers seeking structured domain exploration.

Overall, the summary of our contributions is:

- **Graph-guided critique generation.** We introduce SciCompanion, a framework that combines dynamic multi-hop graph construction, LLM reasoning, and GRPO-optimized retrieval for scientific evaluation.
- **Structured and explainable outputs.** SciCompanion produces feasibility, novelty, and impact critiques with interpretable, evidence-backed reasoning traces.
- **Empirical improvements.** On three peer-review datasets, SciCompanion outperforms RAG and GraphRAG baselines by up to **11.2 points**.
- **Practical utility.** We release an open-source implementation under the MIT License to support reviewers and researchers in critique exploration and literature analysis at SciCompanion[1].

[1]https://github.com/jflashner/SciCompanion

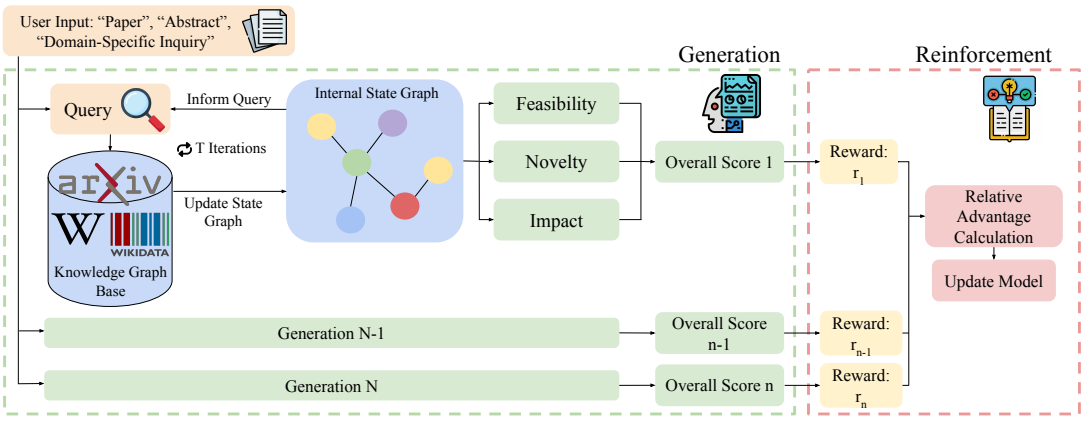

Figure 1: Framework Overview. SCICOMPANION's iterative framework. With T retrieval steps and N GRPO generations, the internal KG is updated, guiding query generation. Final subscores (e.g., feasibility, novelty, impact) inform the overall score.

2 SCICOMPANION: Structure-Aware Reasoning for Scientific Paper Evaluation

Scientific evaluation is a multifaceted task requiring reasoning over text and structured knowledge. SCICOMPANION emulates this expert process using LLMs augmented with structured graph retrieval and RL. Instead of static retrieval or shallow prompting, SCICOMPANION builds a dynamic reasoning system that iteratively constructs context, formulates hypotheses, and aligns judgments with expert evaluations.

2.1 Problem Setup

Given a scientific paper P, the goal is to predict an expert-like assessment vector $\hat{\mathbf{S}} \in \mathbb{R}^k$ covering dimensions such as feasibility, novelty, and impact. Ground-truth labels $\mathbf{Y} \in \mathbb{R}^k$ are sourced from peer-review datasets like PeerRead (Kang et al., 2018) or curated reviews from ICLR and ACL.

To contextualize P, we construct a base scientific knowledge graph $G_{base} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ represent scientific entities and edges $\mathcal{E}$ denote relationships (e.g., citations, derivations, shared methods), built using GraphRAG-style aggregation (Han et al., 2025b). We aim to learn a function f that maps $(P, G_{base}) \mapsto \hat{\mathbf{S}}$ using an LLM-based agent policy π_θ that retrieves relevant evidence, reasons over it, and outputs structured assessments. This policy is optimized using Group Relative Policy Optimization (GRPO), with rewards reflecting both predictive accuracy and reasoning quality (see Section 2.5).

2.2 Framework Overview

SCICOMPANION features three interlinked stages: structured graph retrieval, iterative language-graph reasoning, and multi-dimensional scoring, mimicking expert review. Intuitively, SCICOMPANION 's three stages work synergistically: the first stage retrieves initial "graphlets" of related concepts and references, often incomplete or superficially connected. The second stage iteratively refines these structures by hypothesizing connections, formulating targeted queries, and pruning irrelevant information. Finally, the third stage synthesizes a structured review, explicitly evaluating feasibility, novelty, and impact using the refined graph context. This process is underpinned by two core knowledge representations, a *state graph* G_t (for accumulated structured knowledge) and a *notebook* N_t (for free-form reasoning), evolving jointly as the model queries the KG, updates context, and reflects. The graph structure is crucial: it explicitly represents relational knowledge (capturing dependencies) and supports tractable reasoning over ambiguous or partial knowledge (aiding disambiguation and identification of indirect contributions).

2.3 Structured Retrieval and Graph Completion

Scientific evaluation requires reasoning over explicit content and implicit prior work connections. Static retrieval often fails with specialized terminology, abbreviated references, and assumed domain familiarity, yielding superficial results. Standard RAG's reliance on embedding similarity struggles with semantic depth, especially for dispersed knowledge (Barnett et al., 2024). To address this, SCICOMPANION employs an iterative retrieval-and-reasoning loop, dynamically expanding understanding via structured exploration of a knowledge graph G_{base}.

The process, formalized in Algorithm 1, begins with empty memory structures: a state graph G_0 and a notebook N_0. These two representations, one symbolic, one linguistic, are progressively enriched across T reasoning steps. At each step t, the model generates a new query set Q_t conditioned on the current state (G_{t-1}, N_{t-1}) and the paper P. This conditional formulation ensures that query generation is both context-aware and dynamically tailored, allowing the system to move from broad exploration to focused retrieval as understanding deepens. The queries Q_t are executed over G_{base} to extract a set of subgraphs I_t representing potentially relevant entities, methods, and claims. Retrieved subgraphs I_t are merged into G_{t-1} using symbolic alignment. However, merging alone is insufficient due to scientific expression variability (e.g., synonyms, disconnected facts, implicit relations not in G_{base}). Thus, we introduce *CompGraph*, a policy-driven π_θ graph completion module.

CompGraph proposes edits to the merged graph in three categories: **additions** of novel nodes or edges that reflect claims made in P; **deletions** of outdated or contradicted knowledge; and **revisions** to existing annotations to reflect subtle conceptual shifts. This hybrid symbolic-neural update mechanism ensures that the evolving graph G_t is structurally coherent and semantically aligned with the paper's discourse. The importance of graph completion is twofold. First, it enables the model to reason over latent structure, capturing indirect or compositional contributions that span multiple prior works. Second, it supports robust integration of new information, even when P challenges prevailing knowledge. Notably, this design avoids the need for exhaustive traversal of G_{base}, making reasoning scalable and efficient.

Following the graph update, the model generates an intermediate reasoning trace R_t, appended to the notebook N_t, summarizing its current interpretation of the paper in light of the retrieved and integrated context. The dual memory of graph and notebook supports both explicit symbolic reasoning and flexible abstraction, key properties for emulating expert scientific judgment. After T iterations, the final state (G_T, N_T) captures a structured and context-rich view of the paper's contribution. This state is then passed to a final evaluation module that produces the assessment vector $\hat{\mathbf{S}}$. The full process reflects a balance between structured exploration and reflective synthesis, designed to mimic the expert review process while remaining interpretable and trainable via reinforcement learning (Section 2.5).

2.4 Language-Graph Coupled Reasoning

Scientific evaluation requires more than factual lookup; it demands interpretive reasoning that weighs evidence, identifies assumptions, and contextualizes novelty. Language models without explicit reasoning leave out a crucial planning phase, which helps align generation towards the overall goal. To emulate the reasoning process, SCICOMPANION maintains two complementary representations: a symbolic state graph G_t and a linguistic notebook N_t. At each iteration, the model generates a reasoning trace R_t that reflects its current interpretation of the paper given the retrieved knowledge. This trace is appended to N_t, enabling cumulative, context-aware evaluation. Crucially, this reasoning is not only descriptive but also guides future retrieval. If R_t identifies contradictions or gaps, subsequent queries are adapted accordingly. Over time, the system refines its understanding through this interplay of structured graph (G_t) and reflective reasoning (N_t), yielding a more informed and nuanced evaluation. The final assessment $\hat{\mathbf{S}}$ is produced by analyzing the joint state (G_T, N_T) using dimension-specific prompts. This structured mapping supports interpretability and alignment with expert review criteria. As shown in Figure 1, SCICOMPANION supports multi-sample training: for each paper, we generate multiple reasoning trajectories, each evaluated for scoring accuracy (r^{score}) and structural coherence (r^{struct}). GRPO compares these trajectories to compute relative advantages, updating the policy to favor more coherent and informative reasoning chains. By coupling structured retrieval with iterative reasoning and optimizing for both fidelity and interpretability, SCICOMPANION advances beyond static retrieval systems, offering a transparent and expert-like framework for scientific paper evaluation.

2.5 Policy Optimization via GRPO

To optimize the reasoning and retrieval behaviors in SCICOMPANION, we frame the scientific evaluation task as a reinforcement learning (RL) problem. The agent, parameterized by policy π_θ, is rewarded for generating reasoning trajectories that produce structured evaluations $\hat{\mathbf{S}}$ closely aligned with expert assessments $\mathbf{Y}$. Given the variability in plausible reasoning paths, we adopt Group Relative Policy Optimization (GRPO), which empha-

Algorithm 1 SCICOMPANION Multi-Step Retrieval & Reasoning

Require: Paper P, Base KG G_{base}, Policy π_θ, Steps T
Ensure: Predicted Assessment $\hat{\mathbf{S}}$
Initialize Notebook $N_0 \leftarrow \emptyset$
Initialize State Graph $G_0 \leftarrow \emptyset$
$Q_0 \leftarrow \textit{GenQueries}(\pi_\theta, P, G_0, N_0)$
for $t = 1$ to T **do**
  $I_t \leftarrow \textit{Extract}(G_{base}, Q_{t-1})$
  $G_{merged} \leftarrow \textit{Merge}(G_{t-1}, I_t)$
  $G_t \leftarrow \textit{CompGraph}(\pi_\theta, P, N_{t-1}, G_{merged})$
  $R_t \leftarrow \textit{GenReasoning}(\pi_\theta, P, G_t, N_{t-1})$
  $N_t \leftarrow N_{t-1} \cup \{R_t\}$
  $Q_t \leftarrow \textit{GenQueries}(\pi_\theta, P, G_t, N_t)$
end for
$\hat{\mathbf{S}} \leftarrow \textit{FinalEval}(\pi_\theta, P, G_T, N_T)$
return $\hat{\mathbf{S}}$

sizes relative improvement within a group of candidate responses, promoting exploration without compromising training stability. For each input paper P, we sample N reasoning trajectories using the current policy. Each trajectory produces a predicted score vector $\hat{\mathbf{S}}^{(i)}$ and an associated state graph $G_T^{(i)}$. We then compute two reward components: (1) a score-based reward r_i^{score}, measuring the agreement between $\hat{\mathbf{S}}^{(i)}$ and $\mathbf{Y}$ via RMSE, and (2) a structure-based reward r_i^{struct}, quantifying the informativeness, novelty coverage, and coherence of the final state graph by attribute ratio.

The GRPO objective is given by:

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right) - \beta\, D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right) \tag{1}$$

where $\rho_i = \frac{\pi_\theta(r_i|P)}{\pi_{\theta_{\text{ref}}}(r_i|P)}$ is the importance weight and A_i denotes the relative advantage of trajectory i within its batch.

The training process (Algorithm 2) iteratively samples hypotheses, generates trajectories, computes rewards, and updates the policy via gradient ascent on $J_{\text{GRPO}}(\theta)$. A critical challenge in optimizing SCICOMPANION via GRPO is the inherent complexity of synchronizing query formulation, iterative graph edits, and intermediate reasoning steps. Unlike standard RL scenarios, our task requires sequential, multi-stage token injections within a single forward pass, complicating credit assignment. To address this, we introduce a novel masking technique for GRPO optimization, isolating learning signals specifically to dynamic reasoning actions (queries, graph updates, and reasoning), thereby preventing confounding from static or redundant context. Because the standard GRPO objective assumes the entire generated sequence is produced by the policy, it is not directly applicable to our multi-hop process where search results and prompts are injected externally. Our masking technique resolves this by reformulating the loss function to ensure that only policy-generated tokens contribute to the gradient updates. We define the masked loss for a single trajectory as:

$$\mathcal{L}_{\text{masked}}(\theta) = -\hat{\mathbb{E}}\left[\sum_{i \in \mathcal{M}} \log \pi_\theta(y_i \mid y_{<i}) \rho_i \hat{A}_i\right] \tag{2}$$

where $\mathcal{M}$ is the set of indices corresponding to tokens directly generated by the policy. By constructing the loss in this way, the subsequent gradient calculation, $\nabla_\theta \mathcal{L}_{\text{masked}}(\theta)$, naturally omits the injected content—such as retrieved search results, graph construction prompts, the state graph, and the final review generation prompt—from the optimization. This represents a key technical contribution ensuring stable and mathematically sound optimization for multi-step RL. The use of RL with GRPO allows SCICOMPANION to learn domain-adaptive retrieval and reasoning strategies that generalize across papers and review dimensions, supporting both accuracy and transparency through interpretable outputs.

3 Experiments

To assess the capabilities of SCICOMPANION, we design a comprehensive evaluation protocol grounded in the core challenges outlined in the introduction: scalable critique generation, structured retrieval, and generalization across domains. Our experiments aim to answer three central questions: (Q1) How accurately can SCICOMPANION emulate expert evaluations? (Q2) What is the contribution of multi-step, graph-based retrieval to reasoning quality? (Q3) How does reinforcement learning via GRPO compare to standard prompting and fine-tuning strategies?

Datasets. We use three datasets for evaluation. **ICLR** (5,482 ML papers with reviews) tests multi-dimensional critique (feasibility, novelty, impact) (González-Márquez and Kobak, 2024). **ACL Soundness & Overall** (3,219 CL papers) provide labels for methodological rigor and overall recommendation (Dycke et al., 2025-02). **GoodReads** (5,000 book descriptions with user ratings) tests cross-domain adaptability with loosely structured text and non-expert preferences (Dhamani, 2021).

For all experiments requiring training, we used a standard 80%/10%/10% random split for the training, validation, and test sets, respectively.

Evaluation Metrics. For each dataset, the evaluation task involves predicting a continuous score or vector of scores $\hat{S}$ approximating the expert or crowd-assigned ground truth Y. We report results using three evaluation metrics. *Root Mean Square Error (RMSE)* quantifies predictive accuracy against gold scores. *Point match rates* measure the overlap between generated critiques and peer reviews in terms of strong and weak points. Finally, *retrieval accuracy* is assessed by comparing the system's generated references to ground-truth citations in both full-text and abstract-only settings, providing insight into SciCompanion's ability to surface contextually relevant evidence.

Experimental Setting. Models are tested in zero-shot (guidelines only), five-shot (exemplar reviews), and trained (finetuning and GRPO-based RL) settings, reflecting increasing supervision. Experiments use `GPT-4o-mini`, `Qwen2.5-7B`, and `Qwen2.5-14B` backbones (via `vLLM`, fixed decoding). We used models $\leq$14B to test if our structured evaluation allows them to rival larger unstructured baselines, aiding resource-constrained deployment.

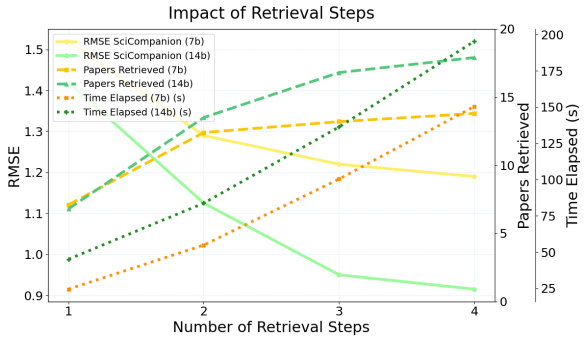

Figure 2: **Impact of retrieval steps:** RMSE and average papers retrieved over iteration steps (K). Multi-step retrieval improves RMSE error by up to 0.5 points.

3.1 Results and Analysis

Effectiveness of Structured Evaluation (Q1). Table 2 presents a comprehensive comparison across zero-shot, five-shot, and trained model settings. Across all datasets, SciCompanion consistently outperforms prompting-only baselines, validating its ability to align with expert judgments through structured, graph-guided reasoning. This trend is further illustrated in Figure 3, which visualizes RMSE across four datasets and three evaluation regimes. In the zero-shot setting, SciCompanion

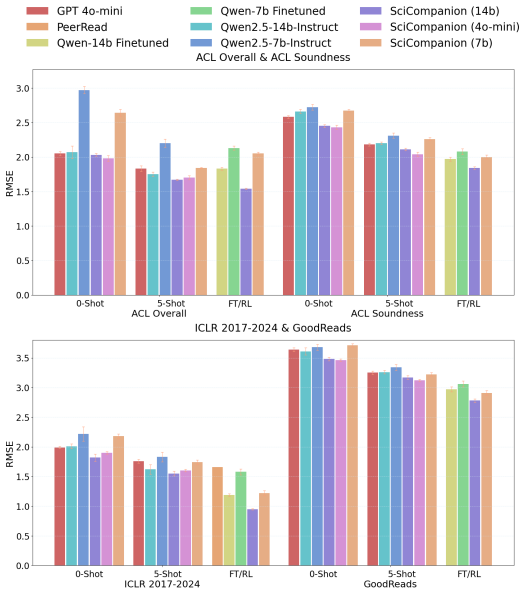

Figure 3: **Lower error rates:** SciCompanion achieves consistently lower RMSE compared to baseline approaches when evaluating scientific hypotheses.

already demonstrates gains over LLMs of comparable size, reducing RMSE by up to *0.33* on ACL Overall and *0.18* on GoodReads. This indicates that even without exemplar reviews, multi-hop retrieval and graph synthesis help surface more relevant contextual evidence. Under five-shot prompting, SciCompanion achieves further reductions, outperforming both baseline and finetuned models. Notably, the improvements persist across both formal peer review datasets (ACL, ICLR) and open-domain corpora (GoodReads), highlighting the generality of our reasoning approach.

The performance advantage becomes most pronounced in the FT/RL regime. On ICLR 2017–2024, SciCompanion with Qwen-14B achieves an RMSE of *0.95*, outperforming the fine-tuned 14B model (RMSE *1.19*) and surpassing PeerRead (RMSE *1.66*). Moreover, our 7B variant of SciCompanion consistently outperforms the 14B prompting baseline across datasets, showcasing that structured critique generation and retrieval alignment can substitute for raw parameter scale. These results confirm SciCompanion's architecture (LLMs, dynamic graph retrieval, GRPO) provides a robust foundation for faithful, interpretable, expert-aligned scientific evaluations.

Ablation Study (Q2).

To isolate the contribution of each core component of SciCompanion, we conducted an ablation study on the ICLR dataset, with results shown in Table 6. Each ablation systematically removes a

Model	RMSE			
	ICLR 2017-2024	ACL Soundness	ACL Overall	GoodReads
Zero-Shot Performance				
GPT 4o-mini	1.99 ± 0.02	2.58 ± 0.02	2.05 ± 0.03	3.64 ± 0.03
Qwen2.5-7b-Instruct	2.22 ± 0.12	2.72 ± 0.04	2.97 ± 0.05	3.68 ± 0.05
Qwen2.5-14b-Instruct	2.01 ± 0.04	2.66 ± 0.03	2.07 ± 0.09	3.61 ± 0.06
SciCompanion (4o-mini)	1.90 ± 0.02	**2.43 ± 0.03**	**1.98 ± 0.04**	**3.46 ± 0.02**
SciCompanion (7b)	2.18 ± 0.04	2.67 ± 0.02	2.64 ± 0.05	3.71 ± 0.03
SciCompanion (14b)	**1.82 ± 0.06**	2.45 ± 0.02	2.03 ± 0.02	3.48 ± 0.03
Five-Shot Performance				
GPT 4o-mini	1.76 ± 0.032	2.18 ± 0.02	1.83 ± 0.04	3.25 ± 0.03
Qwen2.5-7b-Instruct	1.83 ± 0.08	2.31 ± 0.04	2.20 ± 0.06	3.34 ± 0.05
Qwen2.5-14b-Instruct	1.62 ± 0.08	2.20 ± 0.02	1.75 ± 0.03	3.26 ± 0.03
SciCompanion (4o-mini)	1.60 ± 0.02	**2.04 ± 0.03**	1.70 ± 0.03	**3.12 ± 0.02**
SciCompanion (7b)	1.74 ± 0.04	2.26 ± 0.02	1.84 ± 0.01	3.22 ± 0.03
SciCompanion (14b)	**1.55 ± 0.04**	2.11 ± 0.02	**1.67 ± 0.01**	3.17 ± 0.03
Fine-tuned and Reinforcement Learning Models				
Qwen2.5-7b Finetuned	1.58 ± 0.05	2.08 ± 0.04	2.13 ± 0.03	3.06 ± 0.05
Qwen2.5-14b Finetuned	1.19 ± 0.02	1.97 ± 0.03	1.83 ± 0.02	2.97 ± 0.04
PeerRead (2018)	1.66	-	-	-
SEA-EA (2024)	1.31	-	-	-
OpenReviewer	1.72 ±0.05	1.86 ±0.07	1.88 ±0.12	4.83 ±0.27
SciCompanion (7b)	1.22 ± 0.04	2.00 ± 0.03	2.05 ± 0.02	2.91± 0.04
SciCompanion (14b)	**0.95 ± 0.01**	**1.84 ± 0.02**	**1.54 ± 0.01**	**2.78 ± 0.03**

Table 2: Performance comparison across all experimental settings (RMSE) over five runs. Zero-shot describes models prompted only with conference guidelines. Five-shot is provided conference guidelines along with five peer review. The finetuned and reinforcement learning models are provided with the five-shot examples as well as training.

	Weak Match	Strong Match
Gpt-4o-mini	0.322	0.560
Qwen2.5-7B	0.094	0.254
Qwen2.5-14B	0.210	0.394
SciCompanion (7B)	0.370	0.602
SciCompanion (14B)	0.550	0.709

Table 3: Percentage of strong and weak points shared between peer and generated reviews. Examples available in E.1.1

	Retrieval Rate
RAG (Distance)	35.53%
SciCompanion (7B)	38.10%
SciCompanion (14B)	57.50%

Table 4: References Retrieval Rate. Average percentage of references generated matching actual references. Based on the ACL dataset with references ablated.

	Retrieval Rate
RAG (Distance)	27.18%
SciCompanion (7B)	31.10%
SciCompanion (14B)	45.29%

Table 5: Abstract References Retrieval Rate. Average percentage of references generated matching actual references. Based on abstracts from the ACL dataset with references ablated.

Model Configuration	RMSE
SciCompanion (Full Model)	0.95 ± 0.01
(1) w/o CompGraph	1.37 ± 0.03
(2a) w/o GRPO (Fine-tuned only)	1.17 ± 0.02
(2b) w/o GRPO (Prompting only)	1.55 ± 0.03
(3) Single-hop Reasoning Only	1.42 ± 0.04

Table 6: Ablations of SciCompanion (14B) on ICLR 2017-2024. We can see that each component contributes significantly to the prediction accuracy.

key feature from the full model.

Disabling the policy-driven graph completion module ("w/o CompGraph") increases the RMSE from 0.95 to 1.37. This highlights the importance of dynamically completing the knowledge graph to capture latent and implicit connections that are not present in the initial retrieved results.

Next, we evaluated the impact of our training strategy by replacing GRPO with both standard supervised fine-tuning ("w/o GRPO (Fine-tuned only)") and a prompting-only baseline ("w/o GRPO (Prompting only)"). The fine-tuned model's RMSE rises to 1.17, while the prompting-only version performs the worst at 1.55. This wide gap confirms that reinforcement learning via GRPO is critical for aligning the model with the nuanced reward signals of a good critique, significantly outperforming both standard prompting and supervised learning.

Finally, limiting the framework to a single retrieval step ("Single-hop Reasoning Only") degrades performance significantly, with the RMSE increasing to 1.42.

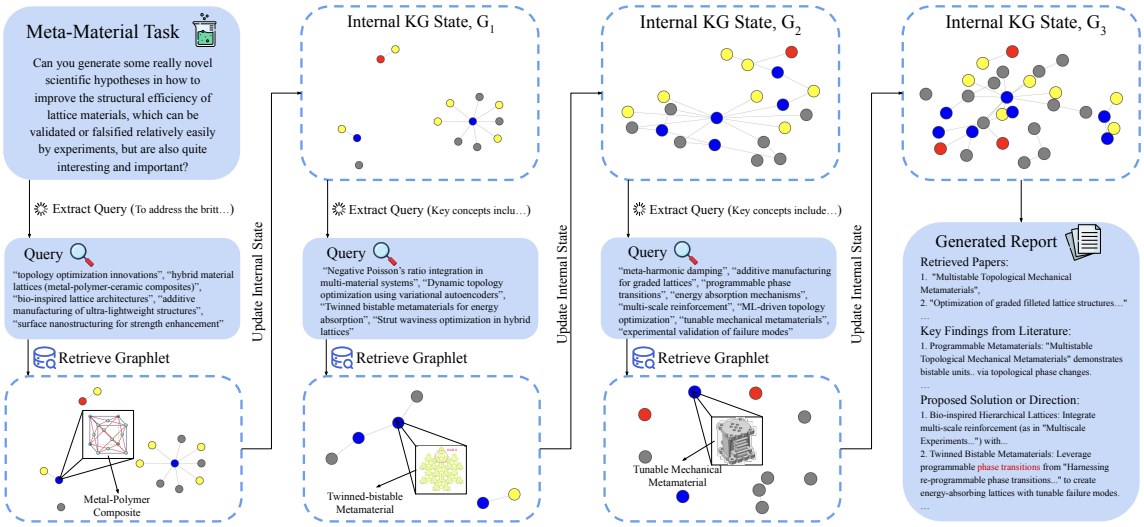

Figure 4: Metamaterial Case Study. We showcase SciCompanion's internal KG update and query generation process using a simple toy example from the field of meta-materials. Pictured is SciCompanion with three iterations (K=3). Arrows represent calls to the LLM for generation. Yellow nodes in the internal knowledge graph represent retrieved papers, red properties, blue materials, and gray methodologies. The shown generated report is a subset of the actual report truncated for demonstration. Material structures adapted from (Yang and Ma, 2020) and (Kappe et al., 2022).

Figure 2 shows a more in-depth analysis revealing that increasing retrieval steps from 1 to 4 steadily decreases RMSE for 7B and 14B models (up to *0.5 points* for 14B). Gains plateau beyond 3 steps (especially for Qwen-14B), suggesting diminishing returns and potential noise from excessive retrieval. We recommend *three retrieval steps* as an optimal trade-off between accuracy and efficiency. Notably, the time elapsed increases nearly linearly with retrieval depth, with the 14B model requiring over 200 seconds at four steps, compared to under 30 seconds at one step. Thus, while deeper retrieval improves reasoning quality, it incurs substantial computational cost. The correlation between papers retrieved and lower RMSE highlights SciCompanion's adaptive querying: unlike static RAG, it dynamically concentrates retrieval on relevant, high-impact literature, improving critique alignment without unnecessary overhead.

Collectively, these ablations demonstrate that each component of SciCompanion; dynamic graph completion, GRPO-based optimization, and multi-hop retrieval, provides a substantial and indispensable contribution to its overall performance.

Review Alignment and Interpretability (Q3). We evaluate how well SciCompanion's generated critiques mirror expert commentary using *point-level match metrics* (Table 3). Both the 7B and 14B variants outperform GPT-4o-mini and Qwen baselines, with SciCompanion (14B) achieving a *70.9% strong point match* rate, over 13 percentage points higher than GPT-4o-mini. These findings indicate that our system is not merely optimizing numerical scores but producing reviews with high conceptual overlap and fidelity. Furthermore, we examine retrieval accuracy as a proxy for evidence-grounding. As shown in Tables 4 and 5, SciCompanion retrieves significantly more ground-truth references than RAG-based models. For instance, in the abstract-only setting, the 14B variant retrieves *45.29%* of actual references, nearly doubling RAG's *27.18%*. This suggests that graph-based iterative retrieval yields more relevant context for critique.

Qualitative Illustration: Metamaterial Case Study. To demonstrate SciCompanion's real-world utility, we include a case study on scientific hypothesis generation in material science (Figure 4). The system iteratively constructs knowledge graphs, generates targeted queries (e.g., "meta-harmonic damping," "topology optimization"), and proposes plausible hypotheses grounded in retrieved literature. The resulting report integrates cross-domain knowledge (e.g., bistable metamaterials, bio-inspired lattices) into structured, testable propositions, mimicking the type of reasoning a domain expert might perform. This example illustrates how SciCompanion's architecture supports *transparent, multi-step discovery*, and highlights its potential for assisting hypothesis refinement and literature exploration. In particular, SciCompanion proposes "leveraging programmable phase transi-

tions... to create energy-absorbing lattices" which is noted to be a feasible and "interesting" research direction by expert evaluators in (Qi et al., 2024).

Summary of Practical Insights. Our experiments reveal several key insights regarding the practical utility of SCICOMPANION. First, *structured reasoning is more critical than scale*, our 7B models outperform prompting-only 14B counterparts, highlighting the value of graph-guided critique generation for small models. Second, *multi-hop retrieval enhances contextual depth*, with three reasoning steps balancing performance and generation time. Finally, *reinforcement learning promotes alignment with expert critiques*, improving both accuracy and interpretability. Together, these results affirm SCICOMPANION as a robust, scalable, and trustworthy scientific assistant capable of supporting peer review and domain exploration workflows.

4 Related Work

AI-assisted scientific discovery and peer review have advanced rapidly, but most systems tackle isolated subtasks (e.g., hypothesis generation, score prediction) rather than structured, critique-aligned evaluation. We categorize related work into scientific discovery and LLM-based evaluation.

AI for Scientific Discovery. AI for scientific discovery has evolved from early expert systems. Modern frameworks like AI Scientist (Lu et al., 2024; Chen et al., 2025) and goal-driven LLM agents (Kumbhar et al., 2025; Zhan et al.) support hypothesis generation but often lack robust validation (feasibility, novelty, impact). RAG (Lewis et al., 2020) and graph-based extensions (GraphRAG (Han et al., 2025b,a), GraphReader (Li et al., 2024)) improve context but can be static, with limited adaptation, and struggle with noisy corpora, retrieval drift, or prioritizing core literature (Barnett et al., 2024).

LLM-Based Evaluation and Peer Review Assistance. PeerRead (Kang et al., 2018) and PEERRec (Bharti et al., 2023) paved the way for LLM-based score prediction, but they rely heavily on surface-level cues like sentiment or style, without modeling deeper scientific structure. Recent interventions (e.g., ICLR 2025 review feedback agents) offer reviewer support but act as prompting aids, not stand-alone evaluators. Domain-specific tools like SciQA (Lu et al., 2022) and SciBench (Wang et al., 2023) target factuality and QA but lack alignment with peer-review dimensions.

Recent works improve review generation via specialized fine-tuning and data processing. For instance, OpenReviewer (Idahl and Ahmadi, 2025) fine-tunes on a large corpus of expert reviews to produce more critical feedback, while PaperSEA (Yu et al., 2024) standardizes inconsistent review data before training. While these methods show the power of data curation and supervised learning, SCICOMPANION differs fundamentally by introducing a dynamic reasoning architecture, leveraging reinforcement learning with structured, graph-grounded reasoning to actively guide critique generation.

Existing methods fall short on: (i) comprehensive claim validation; (ii) multi-hop, graph-structured reasoning; (iii) learning-based retrieval and critique optimization; and (iv) domain adaptability. SCICOMPANION addresses these limitations through dynamic graph construction, critique-aligned reasoning, and GRPO-based self-improvement.

5 Conclusion

We present SCICOMPANION, a critique-aligned, graph-grounded reasoning framework for structured scientific evaluation. Motivated by the rising scale and complexity of peer review, SCICOMPANION combines large language models with dynamic knowledge graphs and reinforcement learning to perform transparent, multi-hop assessments of scientific work. Its architecture reflects how expert reviewers navigate literature, retrieving relevant prior work, reasoning over structured evidence, and grounding judgments in contextual understanding. Our experiments across four diverse datasets demonstrate that SCICOMPANION substantially improves evaluation quality, reducing RMSE by up to 31.2% compared to prompting-only baselines. Through structured graph construction and GRPO-based optimization, the framework enables smaller models (e.g., 7B) to match or exceed the performance of larger, unstructured counterparts, offering a practical, scalable solution for review assistance and domain exploration. By aligning LLM behavior with scientific critique dimensions (feasibility, novelty, impact), SCICOMPANION advances the frontier of trustworthy, interpretable AI for science. It offers a reproducible, extensible approach to enhance peer review and paves the way for future systems supporting hypothesis generation, literature synthesis, and human-AI discovery.

6 Limitations

While SciCompanion demonstrates strong empirical performance and interpretability, several limitations remain. First, its reliance on curated knowledge graphs and open-access corpora may restrict coverage in underrepresented or rapidly evolving scientific domains. As a result, evaluation quality may degrade when source graphs are sparse or incomplete. Second, although our GRPO optimization improves alignment with expert assessments, it requires supervised review data that may not be available in all disciplines. Third, our evaluation primarily focuses on English-language scientific texts; the framework's generalizability to multilingual or low-resource scientific communities remains untested. Additionally, while point-matching metrics capture surface agreement with human reviews, they do not fully reflect deeper aspects of critique quality, such as originality, fairness, or epistemic humility. Finally, we do not yet evaluate the long-term effects of automated review assistance on human decision-making or reviewer behavior, which would be important for safe deployment in academic peer review pipelines.

7 Ethics Statement

This work aims to assist scientific evaluation through structured reasoning and knowledge-grounded critique generation. All datasets used are publicly available and derived from peer-reviewed or crowd-sourced domains (e.g., ICLR, ACL, GoodReads), and do not contain personal or sensitive information. No human subjects were involved in data collection, annotation, or evaluation. We acknowledge that automated assessments may inadvertently reinforce existing biases in peer review datasets or favor dominant scientific paradigms. SciCompanion is not intended to replace expert judgment but to augment human reviewers with transparent, evidence-backed reasoning. We strongly recommend that any use of this system in high-stakes review or discovery contexts involve human oversight and be accompanied by explanations and uncertainty estimates. Our design emphasizes interpretability and critique alignment to mitigate risks of overreliance on opaque model predictions. Nonetheless, further work is needed to ensure fairness, accountability, and inclusivity in AI-assisted scientific evaluation.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported by the National Science Foundation under Award No. IIS-2339989 and No. 2406439, DARPA under contract No. HR00112490370 and No. HR001124S0013, U.S. Department of Homeland Security under Grant Award No. 17STCIN00001-08-00, Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, Amazon AWS, Google, Cisco, 4-VA, Commonwealth Cyber Initiative, National Surface Transportation Safety Center for Excellence, and Virginia Tech. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

Rakeshwar Bandichhor, AS Borovik, Ana de Bettencourt-Dias, Martin D Eastgate, Nora S Radu, Feng Shi, and Lisa McElwee-White. 2023. In support of early-career researchers.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *Preprint*, arXiv:2401.05856.

Prabhat Kumar Bharti, Tirthankar Ghoshal, Mayank Agarwal, and Asif Ekbal. 2023. Peerrec: An ai-based approach to automatically generate recommendations and predict decisions in peer review. *International Journal on Digital Libraries*, 25:55–72.

Jianpeng Chen, Wangzhi Zhan, Haohui Wang, Zian Jia, Jingru Gan, Junkai Zhang, Jingyuan Qi, Tingwei Chen, Lifu Huang, Muhao Chen, and 1 others. 2025. Metamatbench: Integrating heterogeneous data, computational tools, and visual interface for metamaterial discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5334–5344.

DeepSeek-AI and et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, Davide Buscaldi, and Enrico Motta. 2021. Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. *Future Generation Computer Systems*, 116:253–264.

Manav Dhamani. 2021. Goodreads 100k books.

Nils Dycke, Lu Sheng, Hanna Holtdirk, and Iryna Gurevych. 2025-02. Nlpeerv2: A unified resource for the computational study of peer review.

Rita González-Márquez and Dmitry Kobak. 2024. Learning representations of learning representations. In *Data-centric Machine Learning Research (DMLR) workshop at ICLR 2024*.

Salman Y Guraya, Robert I Norman, Khalid I Khoshhal, Shaista Salman Guraya, and Antonello Forgione. 2016. Publish or perish mantra in the medical field: A systematic review of the reasons, consequences and remedies. *Pakistan journal of medical sciences*, 32(6):1562.

Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025a. Rag vs. graphrag: A systematic evaluation and key insights. *Preprint*, arXiv:2502.11371.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025b. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309*.

Maximilian Idahl and Zahra Ahmadi. 2025. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *Preprint*, arXiv:2412.11948.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Rachelle W Johnson and Megan M Weivoda. 2021. Current challenges for early career researchers in academic research careers: Covid-19 and beyond.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies (Volume 1: Long Papers)*, pages 1647–1661. Association for Computational Linguistics.

Konstantin Kappe, Jan P Wahl, Florian Gutmann, Silviya M Boyadzhieva, Klaus Hoschke, and Sarah C L Fischer. 2022. Design and manufacturing of a metal-based mechanical metamaterial with tunable damping properties. *Materials (Basel)*, 15(16):5644.

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. 2025. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents. *arXiv preprint arXiv:2501.13299*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, Wenbo Su, and Bo Zheng. 2024. Graphreader: Building graph-based agents to enhance long-context abilities of large language models. *arXiv preprint arXiv:2406.14550*.

Yicheng Liu. 2025. Retrieval-augmented generation: Methods, applications and challenges. *Applied and Computational Engineering*, 142:99–108.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 2503–2516.

Bahar Mehmani and Ashutosh Ghildiyal. 2024. Rethinking reviewer fatigue. *Editorial Office News*, 17.

Allard Oelen, Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Generate fair literature surveys with scholarly knowledge graphs. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 97–106.

Eugenio Picano. 2025. Who is a reviewer? the good, the bad, and the ugly phenotypes.

Jingyuan Qi, Zian Jia, Minqian Liu, Wangzhi Zhan, Junkai Zhang, Xiaofei Wen, Jingru Gan, Jianpeng Chen, Qin Liu, Mingyu Derek Ma, Bangzheng Li, Haohui Wang, Adithya Kulkarni, Muhao Chen, Dawei Zhou, Ling Li, Wei Wang, and Lifu Huang. 2024. Metascientist: A human-ai synergistic framework for automated mechanical metamaterial design. *Preprint*, arXiv:2412.16270.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

H. Xiong, S. Li, Y. Feng, Z. Yang, Z. Liu, and M. Sun. 2024. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*.

Yixuan Even Xu, Fei Fang, Jakub Tomczak, Cheng Zhang, Zhenyu Sherry Xue, Ulrich Paquet, and Danielle Belgrave. 2024. Neurips 2024 experiment on improving the paper-reviewer assignment.

Hang Yang and Li Ma. 2020. Angle-dependent transitions between structural bistability and multistability. *Advanced Engineering Materials*, 22.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *Preprint*, arXiv:2407.12857.

Xinyue Zeng, Haohui Wang, Junhong Lin, Jun Wu, Tyler Cody, and Dawei Zhou. Lensllm: Unveiling fine-tuning dynamics for llm selection. In *Forty-second International Conference on Machine Learning*.

Wangzhi Zhan, Jianpeng Chen, Dongqi Fu, and Dawei Zhou. Unimate: A unified model for mechanical metamaterial generation, property prediction, and condition confirmation. In *Forty-second International Conference on Machine Learning*.

A Notation

Table 7: Key Symbols and Definitions

Symbol	Definition
P	Scientific paper evaluated
G_{base}	Base knowledge graph
G_t	State graph at step t
$\mathbf{Y}$	Ground truth expert assessment vector
$\hat{\mathbf{S}}$	Predicted assessment vector by model
π_θ	Agent policy (parameterized by θ)
Q_t	Queries generated at step t
I_t	Information extracted at step t
R_t	Intermediate reasoning at step t
N_t	Notebook state at step t
r	RL reward signal
J_{GRPO}	GRPO objective function

B Hyperparameters and Settings

This section details the hyperparameters and settings used for the Qwen2.5 7B and 14B models, including model loading with LoRA, inference generation, and GRPO training (Qwen et al., 2025). Training was conducted on $4\times$ A100 GPUs. The total computational budget for all experiments is estimated to be approximately 340 GPU hours on this hardware.

B.1 Model Loading and LoRA Configuration (Qwen2.5 7B & 14B)

The Qwen2.5 7B and 14B models were loaded using the Unsloth library's `FastLanguageModel`. Key settings for loading the base model and configuring PEFT (LoRA) are listed below.

B.1.1 Base Model Loading (`FastLanguageModel.from_pretrained`)

- `model_name`: "unsloth/Qwen2.5-7B-Instruct" or "unsloth/Qwen2.5-14B-Instruct"
- `dtype`: `torch.bfloat16`
- `load_in_4bit`: `True`
- `fast_inference`: `True`
- `gpu_memory_utilization`: `0.4`
- `max_seq_length`: `24000`
- `max_lora_rank`: `128` (matches `lora_rank`)

B.1.2 PEFT Model Configuration (`FastLanguageModel.get_peft_model`)

- LoRA Rank (`r`): `128`
- Target Modules (`target_modules`):
  - `"q_proj"`
  - `"k_proj"`
  - `"v_proj"`
  - `"o_proj"`

  - "gate_proj"
  - "up_proj"
  - "down_proj"
- LoRA Alpha (lora_alpha): 256 (calculated as $2 \times$ lora_rank)

B.2 Generation Hyperparameters (Inference)

The following settings from GenerationConfig were used during inference:

- num_return_sequences: 1
- max_new_tokens: 4800
- temperature: 0.6
- top_p: 0.95
- top_k: 20
- do_sample: True

B.3 GRPO (Group Relative Policy Optimization) Settings

The GRPOConfig was used for training with the following parameters:

- use_vllm: True
- learning_rate: 1e-5
- adam_beta1: 0.9
- adam_beta2: 0.99
- weight_decay: 0.1
- warmup_ratio: 0.1
- temperature (for GRPO policy sampling): 1.0
- lr_scheduler_type: "cosine"
- optim: "adamw_8bit"
- bf16: True
- gradient_accumulation_steps: 1
- num_generations (for GRPO): 8
- max_prompt_length: 12000
- max_completion_length: 2048
- num_train_epochs: 30
- max_steps: 300
- save_steps: 300
- max_grad_norm: 0.2

C Algorithms

C.1 GRPO Reinforcement

This algorithm trains a retrieval and reasoning policy π_θ using *Group Relative Policy Optimization (GRPO)* to align model-generated evaluations with expert judgments. The policy parameters are initialized as $\theta = \theta_0$. In each training iteration, a batch of hypotheses $\{H_j\}$ is sampled, and the current policy π_θ generates reasoning trajectories over them. Rewards are computed based on evaluation accuracy, typically reflecting alignment with expert-assigned scores. The policy is then updated using the GRPO objective via gradient ascent: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$. This iterative process enables the policy to learn adaptive retrieval and critique behaviors that generalize across domains. The optimized policy π_θ is returned upon completion.

Algorithm 2 Retrieval Policy Optimization

Require: Training dataset of hypotheses and expert
Ensure: Optimized policy π_θ
Initialize policy parameters $\theta = \theta_0$
for each training iteration **do**
  Sample batch of hypotheses $\{H_j\}$
  Collect trajectories using current policy
  Compute rewards based on evaluation accuracy
  Update policy using GRPO:
  $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
end for
return π_θ

D Licensing and Intended Use

This research adheres to the licensing terms of all artifacts used and created.

D.1 SciCompanion Artifact

The open-source implementation of SciCompanion, including all code associated with this paper, is released under the MIT License.

D.2 Datasets

Our use of the following datasets for academic research and model training is consistent with their intended purpose and licensing terms:

- The ICLR dataset (González-Márquez and Kobak, 2024) is provided under an MIT license.

- The ACL Soundness & Overall dataset (Dycke et al., 2025-02) is provided under a CC-BY-NC-4.0 license, which permits non-commercial research use. Our work fully complies with this non-commercial stipulation.
- The GoodReads dataset is available under a CC0: Public Domain license.

Our framework is intended for research purposes to assist in scientific evaluation and is not deployed for commercial use.

E Point Matching

E.1 Points Matches

	Weak Match	Strong Match
Gpt-4o-mini	0.322	0.560
Qwen2.5-7B	0.094	0.254
Qwen2.5-14B	0.210	0.394
SciCompanion (7B)	0.370	0.602
SciCompanion (14B)	0.550	0.709

Table 8: Percentage of strong and weak points shared between peer and generated reviews. Calculated as the number of common points over the total number of weak and strong comments in the peer review, respectively.

Table 8 reports the percentage of weak and strong review points generated by each model that align with corresponding peer reviewer comments. SciCompanion significantly outperforms all baselines across both weak and strong point matches. Notably, the 14B variant achieves a **strong match rate of 70.9%** and a **weak match rate of 55.0%**, indicating high fidelity to expert evaluations. Even the 7B version surpasses GPT-4o-mini, achieving higher alignment despite using fewer parameters. In contrast, standard models like Qwen2.5-7B and Qwen2.5-14B exhibit considerably lower match rates, particularly on weak points, suggesting less interpretive depth and alignment. These results demonstrate that SciCompanion produces critiques that are not only accurate in score but also substantively consistent with human reviewers in terms of both strengths and weaknesses.

E.1.1 Examples

Figures 5–10 present qualitative comparisons between peer reviews and generated reviews for three representative papers. Across samples, we observe that SciCompanion not only replicates key strong points, such as addressing bias in healthcare (Figure 5) or identifying the novelty of a method (Figure 7), but often provides more comprehensive justifications. Similarly, in weak point comparisons (Figures 6, 8, 10), the model highlights limitations related to dataset scope, technical clarity, and generalizability that closely mirror expert concerns. In several cases, the generated critiques go further by suggesting concrete improvements or clarifying implications. These examples underscore the model's ability to emulate expert reasoning at a fine-grained level, reinforcing the point-level match metrics with substantive evidence of interpretive depth and contextual relevance.

Peer Review Strong Points (2)	Generated Review Strong Points (5)
The paper tackles the critical issue of bias in LLMs within the healthcare domain, a vital area given the potential for impact on real-world health outcomes.	The paper introduces a systematic approach to evaluate biases in LLMs by curating a dataset of context-dependent healthcare questions and comparing model responses with and without demographic context. (match score: 0.60)
The experimental design incorporates a broad range of demographic variables (age, sex, location), enhancing the relevance and applicability of the findings across diverse populations.	The study provides empirical evidence on biases in LLM responses across different demographic attributes, contributing to the broader understanding of how LLMs can be biased in healthcare applications.
	The paper is well-organized and effectively summarizes relevant prior work, providing a strong foundation for the study's contributions.
	The insights from this study are relevant for both broader audiences interested in LLMs and narrower audiences focused on healthcare applications, making it a valuable contribution to the field.
	The study is significant in understanding and mitigating biases in LLMs, especially within the healthcare domain, where accurate and unbiased responses are crucial. (match score: 0.85)

Figure 5: Sample 15 Strong Points. Both the peer review and generated review note the tackling of bias in the healthcare domain as a significant strong point. To a lesser extent, both reviews mention demographic context as a strong point.

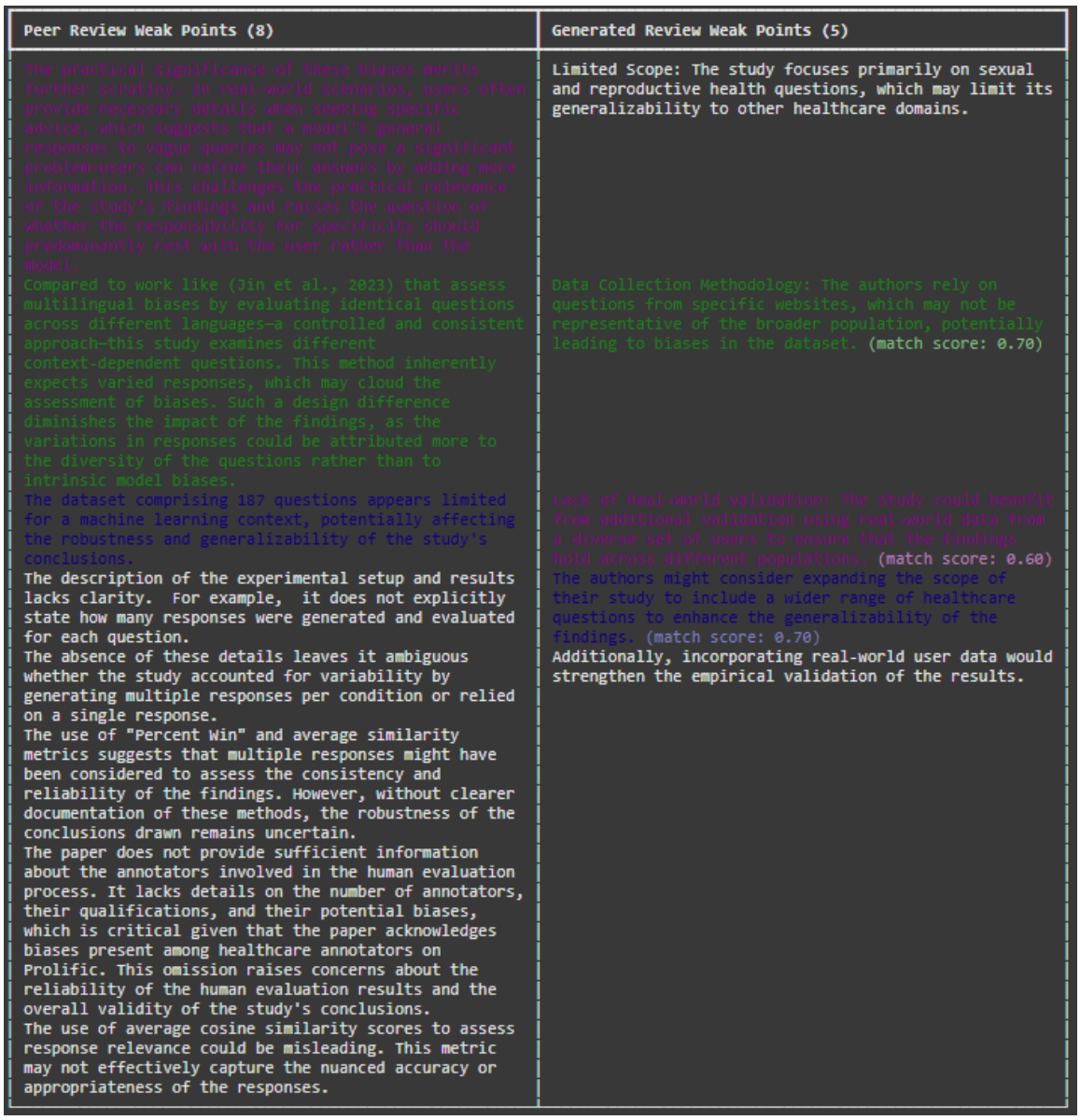

Figure 6: Sample 15 Weak Points. Both reviews question the real-world relevance of the study's findings. Both reviews also point out that the observed bias could be from the source of the questions rather than the model itself. Both reviews recommend widening the dataset to enhance generalizability.

Peer Review Strong Points (2)	Generated Review Strong Points (12)
This paper proposes a new trainable method that combines transformer-based LMs and GNN.	The paper introduces a novel approach to enhance the accuracy of factual reasoning in knowledge graph question answering (KGQA) tasks by minimizing the impact of noisy pathways in retrieved subgraphs.
Experimental results show that the proposed method can effectively inject knowledge and achieve significant improvements.	The proposed method, Q-KGR, aims to eliminate irrelevant information from knowledge graphs by assigning relevance scores to edges based on the input question.
	This method is further complemented by the introduction of Knowformer, a custom transformer, which aligns and injects structured knowledge into the parameter space of Large Language Models (LLMs) to enhance their factual reasoning capabilities. (match score: 0.60)
	The paper builds on existing research in KGQA and addresses a significant challenge in the field, namely the presence of noisy pathways that can impede the model's ability to perform accurate reasoning.
	The proposed method shows promising results across multiple KGQA datasets, demonstrating significant improvements over existing systems.
	This paper introduces a novel approach, Q-KGR (Question-guided Knowledge Graph Re-scoring), to enhance factual reasoning in Knowledge Graph Question Answering (KGQA) tasks by eliminating noisy pathways.
	Additionally, the paper introduces Knowformer, a custom transformer that injects structured knowledge into LLMs to improve their factual reasoning capabilities. (match score: 0.60)
	The proposed methods are evaluated on several KGQA benchmarks, demonstrating significant improvements over existing systems.
	The paper presents a novel approach, Q-KGR, which addresses a core challenge in KGQA by eliminating noisy pathways. This is a significant contribution to the field.
	The evaluation results on multiple KGQA benchmarks (Open BookQA, AI2 Reasoning Challenge, RiddleSense, and Physical Interaction Question Answering) demonstrate the effectiveness of the proposed method.
	The paper provides a solid theoretical foundation for both Q-KGR and Knowformer, which enhances the credibility of the proposed solutions.
	The paper is well-organized and clearly presents the methodology, experimental setup, and results.

Figure 7: Sample 18 Strong Points. In this sample, we see that the generated review is far more comprehensive in its coverage than the corresponding peer review. Both reviews identify the source of novelty in the paper as well as the core contribution of the method.

Peer Review Weak Points (4)	Generated Review Weak Points (4)
Lacking compared methods.	The paper could benefit from more detailed real-world validation, as the current evaluations are primarily based on synthetic datasets. (match score: 0.60)
The used datasets are more like knowledge-sensitive common-sense datasets rather than typical *question answering over knowledge graph* (KGQA) datasets that *finds answers directly from the knowledge graph*. Consequently, KAPING, a KGQA method, might not be a suitable baseline.	The proposed methods (Q-KGR and Knowformer) are complex and might require a significant amount of resources to implement, which could limit their practical applicability.
The applicability of the re-scoring module needs to be discussed.	The paper could benefit from additional discussion on the potential scalability of the proposed methods to larger knowledge graphs. (match score: 0.60)
Some key technical details may need further clarification.	Providing more detailed insights into the alignment process between the structured knowledge and LLMs would be beneficial. (match score: 0.60)

Figure 8: Sample 18 Weak Points. Both reviews highlight the need for more challenging QA datasets. Both reviews question how applicable the method is to knowledge graphs not evaluated in the paper. Both reviews ask for clarification of technical details.

Peer Review Strong Points (2)	Generated Review Strong Points (7)
Demonstrated high performance in benchmarks	MSI-Agent introduces a novel method for efficiently summarizing and utilizing multi-scale insights, addressing the challenges of irrelevant and lacking general insights in decision-making processes for embodied agents.
Demonstrated robustness to domain shift	The paper provides a clear and detailed description of the methodology, including the three-part pipeline (experience collection, insight generation, and insight utilization). The approach is well-structured and offers a comprehensive framework for insight utilization in embodied agents.
	The authors present a thorough empirical evaluation, showing significant improvements in task success rates compared to alternative methods. These results are compelling and support the effectiveness of the approach. (match score: 0.60)
	The work is highly relevant to the field of embodied AI and LLMs for task planning, addressing real-world challenges in decision-making and task execution.
	This study provides sufficient support for all of its claims.
	Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.
	I think this paper could be accepted to an *ACL conference.

Figure 9: Sample 23 Strong Points. We again see that the generated review is longer and more comprehensive than the peer review. Both reviews mention the empirical results as a strong point of the paper.

Peer Review Weak Points (5)	Generated Review Weak Points (9)
The authors claim that the proposed method consists of multiple modules and that their integration enables efficient use of insights, and I'd like to see an ablation study.	The approach involves multiple steps and components, which might introduce complexity in implementation and practical deployment. (match score: 0.60)
By measuring the effects of success mode/pair mode in Experience Selection, the effects of dividing Multi-Scale Insight into General/subtask, and the effects of Hashmap indexing/vector indexing in Multi-Scale Insight Selection, it can be clarified which mechanisms contributed more to performance improvement.	Clear guidelines and tools for implementation would be beneficial.
There is no detailed validation regarding whether the insights selected by the proposed method are truly task-specific, high-level, and free of irrelevant information, as claimed by the authors.	While the approach shows promise in the specific task environments tested, its generalization to a broader range of tasks and domains is not fully demonstrated. (match score: 0.60)
Figure 4 only shows one example of the difference between Expel Insight Memory and MSE Insight Memory.	Further experiments on diverse tasks would strengthen the paper's claims.
If the authors are asserting the refinement of insights by the proposed method, I'd like to see a more thorough comparative validation of the insights produced by both methods	The comparisons to baseline methods are robust, but the impact of the method's success-failure experience pairs versus success-only experience pairs on task success rates should be more thoroughly explored and justified. (match score: 0.65)
	The paper could benefit from a more detailed discussion on how the multi-scale insights are generated and how they are utilized during task execution.
	Providing more specific examples or visual aids would clarify these processes.
	Consideration of potential challenges in deploying the approach in real-world settings, particularly in terms of computational requirements and data collection, would enhance the practical value of the work.
	The paper's use of large language models (LLMs) could be expanded to discuss their limitations and how they contribute to the approach, providing a more comprehensive analysis.

Figure 10: Sample 23 Weak Points. Both reviews touch on the complexity of the method as a weak point. Both reviews question the validation of the method on a wider range of tasks. Both reviews request a more rigorous comparison between the insights produced by both methods in the paper.

E.2 References Recreation

E.2.1 From Paper

	Retrieval Rate
RAG (Distance)	35.53%
SciCompanion (7B)	38.10%
SciCompanion (14B)	57.50%

Table 9: References Retrieval Rate. Calculated as the average number of references generated out of ten matching actual references. Results based on papers from the ACL ARR dataset with references removed.

Table 9 compares the reference retrieval accuracy of SCICOMPANION against a standard distance-based RAG baseline. The 14B variant of SCICOMPANION achieves a retrieval rate of **57.5%**, substantially outperforming both the RAG baseline (35.53%) and its 7B counterpart (38.10%). This indicates that the model's graph-guided, multi-hop retrieval mechanism is more effective at surfacing contextually relevant citations from the scientific literature. The improvement is particularly notable given that the ACL ARR dataset used in this evaluation has references ablated, making the task more challenging. These results highlight the framework's ability to reconstruct meaningful scholarly context, which is essential for producing grounded and trustworthy scientific evaluations.

E.2.2 Example Reasoning Trace:

```
[[ ## reasoning ## ]]
The review focuses on the paper's contributions to
↪  causal interpretability in multimodal models,
↪  its methodological rigor, and alignment with
↪  existing literature. The related papers list
↪  emphasizes works on causal mediation,
↪  adversarial learning, and multimodal
↪  benchmarks, which are central to the paper's
↪  claims. The review evaluates the paper's
↪  novelty, methodology, and potential impact
↪  while addressing concerns about scalability,
↪  dataset bias, and causal validity.

[[ ## related_papers ## ]]
[
  "CausaLM: Causal Model Explanation Through
  ↪  Counterfactual Language Models",
  "Amnesic Probing: Removing Specific Information
  ↪  for Model Interpretability",
  "NLX-GPT: Adversarial Learning for Text
  ↪  Generation Attribution",
  "Causal Inference with Deep Causal Graphs",
  "Explaining the Behavior of Black-Box Prediction
  ↪  Algorithms with Causal Learning",
  "Deep Structural Causal Models for Tractable
  ↪  Counterfactual Inference",
  "Interpreting Deep Models through the Lens of
  ↪  Data",
  "Causal Inference in Natural Language
  ↪  Processing: Estimation, Prediction,
  ↪  Interpretation and Beyond",
  "Generative causal explanations of black-box
  ↪  classifiers",
  "Learning Latent Causal Structures with a
  ↪  Redundant Input Neural Network"
]

[[ ## review ## ]]
**Paper Summary**
This paper introduces a causal framework for
↪  interpreting VisualBERT predictions in
↪  offensive meme detection, leveraging
↪  structural causal models (SCMs),
↪  de-confounding, adversarial learning, and
↪  dynamic routing. It addresses limitations of
↪  input attribution methods (e.g., Integrated
↪  Gradients, Saliency) by emphasizing causal
↪  reasoning and evaluates model trustworthiness
↪  through ablation studies. The framework is
↪  tested on the Facebook Hateful Meme Dataset, a
↪  benchmark for multimodal content moderation.

**Summary of Strengths**
1. **Novel Methodology**: Proposes a causal
↪  framework integrating SCMs with multimodal
↪  models (VisualBERT), addressing non-causal
↪  attributions and implicit content in
↪  offensive memes.
2. **Empirical Rigor**: Quantitatively evaluates
↪  causality through ablation studies and
↪  compares with prior methods (e.g., CausaLM,
↪  Amnesic Probing).
3. **Relevance to Safety-Critical Applications**:
↪  Highlights the importance of causal
↪  reasoning in deploying trustworthy AI for
↪  content moderation, aligning with trends in
↪  explainable AI (e.g., "Explainable Deep
↪  Learning: A Field Guide for the Uninitiated").
```

```
4. **Technical Contribution**: Introduces
↪  de-confounding and dynamic routing as
↪  scalable solutions for local causal
↪  interpretability, building on prior work
↪  like "Deep Structural Causal Models for
↪  Tractable Counterfactual Inference."

**Summary of Weaknesses**
1. **Limited Dataset Scope**: Relies on the
↪  Facebook Hateful Meme Dataset, which may not
↪  generalize to other domains or languages.
2. **Causal Validity Concerns**: The framework
↪  assumes linear relationships between
↪  concepts and latent representations, which
↪  may oversimplify complex interactions (e.g.,
↪  "Causal Inference in Natural Language
↪  Processing" critiques such assumptions).
3. **Scalability**: Dynamic routing and
↪  adversarial learning may increase
↪  computational costs, limiting practical
↪  deployment (similar to challenges noted in
↪  "Iterative Nullspace Projection" for
↪  large-scale models).
4. **Bias in Concept Annotation**: The annotated
↪  concept set (e.g., "racism," "holocaust")
↪  introduces potential subjectivity, as
↪  discussed in "Interpreting Deep Models
↪  through the Lens of Data."

**Comments/Suggestions/Typos**
- Clarify how the framework handles concept
↪  overlaps (e.g., "Holocaust" and "Nazism" in
↪  Table 1).
- Expand evaluation to other datasets (e.g.,
↪  Multimodal Sentiment Analysis benchmarks) to
↪  demonstrate robustness.
- Address potential biases in concept annotation
↪  by incorporating inter-annotator agreement
↪  metrics.

**Reviewer Confidence**
4 = Quite sure: The paper's contributions are
↪  well-defined, and methodology aligns with
↪  existing literature on causal
↪  interpretability.

**Soundness**
4 = Strong: The framework is technically sound,
↪  with clear claims supported by ablation
↪  studies. However, the assumption of linearity
↪  in concept-latent interactions warrants
↪  further validation.

**Excitement**
3.5 = Interesting: The work is relevant to
↪  safety-critical AI and advances causal
↪  interpretability, but its impact may be
↪  domain-specific.

**Overall Assessment**
4 = Conference: The paper makes a clear
↪  contribution to causal interpretability in
↪  multimodal models and is suitable for
↪  publication in a selective *ACL venue.

**Best Paper Justification**
N/A (Overall assessment is "Conference," not
↪  "Consider for Award").

[[ ## overall_score ## ]]
4.0
[[ ## completed ## ]]
```

F Prompts

F.1 ACL Prompt for Final Score Generation:

```
Your input fields are:
       1. `paper` (str): Computer science paper to review

Your output fields are:
       1. `review` (str):
           Review Form
               Paper Summary
              Describe what this paper is about. This should help action
              ↪  editors and area chairs to understand the topic of the
              ↪  work and highlight any possible misunderstandings.

               Summary of Strengths
               What are the major reasons to publish this paper at a
              ↪  selective *ACL venue? These could include novel and
              ↪  useful methodology, insightful empirical results or
              ↪  theoretical analysis, clear organization of related
              ↪  literature, or any other reason why interested
              ↪  readers of *ACL papers may find the paper useful.

               Summary of Weaknesses
               What are the concerns that you have about the paper that
              ↪  would cause you to favor prioritizing other
              ↪  high-quality papers that are also under consideration
              ↪  for publication? These could include concerns about
              ↪  correctness of the results or argumentation, limited
              ↪  perceived impact of the methods or findings (note
              ↪  that impact can be significant both in broad or in
              ↪  narrow sub-fields), lack of clarity in exposition, or
              ↪  any other reason why interested readers of *ACL
              ↪  papers may gain less from this paper than they would
              ↪  from other papers under consideration. Where
              ↪  possible, please number your concerns so authors may
              ↪  respond to them individually.

               Comments/Suggestions/Typos
              If you have any comments to the authors about how they may
              ↪  improve their paper, other than addressing the
              ↪  concerns above, please list them here.

               Reviewer Confidence
             5 = Positive that my evaluation is correct. I read the paper
              ↪  very carefully and am familiar with related work.
              4 = Quite sure. I tried to check the important points
              ↪  carefully. It's unlikely, though conceivable, that I
              ↪  missed something that should affect my ratings.
              3 = Pretty sure, but there's a chance I missed something.
              ↪  Although I have a good feel for this area in general,
              ↪  I did not carefully check the paper's details, e.g.,
              ↪  the math or experimental design.
            2 = Willing to defend my evaluation, but it is fairly likely that
              ↪  I missed some details, didn't understand some central
              ↪  points, or can't be sure about the novelty of the work.
              1 = Not my area, or paper is very hard to understand. My
              ↪  evaluation is just an educated guess.
               Soundness
             Given that this is a long paper, is it sufficiently sound and
              ↪  thorough? Does it clearly state scientific claims and
              ↪  provide adequate support for them? For experimental
              ↪  papers: consider the depth and/or breadth of the
              ↪  research questions investigated, technical soundness
              ↪  of experiments, methodological validity of evaluation.
              ↪  For position papers, surveys: consider whether the
              ↪  current state of the field is adequately represented
              ↪  and main counter-arguments acknowledged. For resource
              ↪  papers: consider the data collection methodology,
              ↪  resulting data & the difference from existing
              ↪  resources are described in sufficient detail.

               5 = Excellent: This study is one of the most thorough I
              ↪  have seen, given its type.
               4.5
              4 = Strong: This study provides sufficient support for all
              ↪  of its claims. Some extra experiments could be nice,
              ↪  but not essential.
               3.5
              3 = Acceptable: This study provides sufficient support for
              ↪  its main claims. Some minor points may need extra
              ↪  support or details.
               2.5
            2 = Poor: Some of the main claims are not sufficiently supported.
              ↪  There are major technical/methodological problems.
               1.5
             1 = Major Issues: This study is not yet sufficiently thorough
              ↪  to warrant publication or is not relevant to ACL.
               Excitement
```

```
        How exciting is this paper for you? Excitement is
       ↪  subjective, and does not necessarily follow what is
       ↪  popular in the field. We may perceive papers as
       ↪  transformational/innovative/surprising, e.g. because
       ↪  they present conceptual breakthroughs or evidence
       ↪  challenging common
       ↪  assumptions/methods/datasets/metrics. We may be
       ↪  excited about the possible impact of the paper on
       ↪  some community (not necessarily large or our own),
       ↪  e.g. lowering barriers, reducing costs, enabling new
       ↪  applications. We may be excited for papers that are
       ↪  relevant, inspiring, or useful for our own research.
       ↪  These factors may combine in different ways for
       ↪  different reviewers.

      5 = Highly Exciting: I would recommend this paper to others
       ↪  and/or attend its presentation in a conference.
       4.5
      4 = Exciting: I would mention this paper to others and/or make
       ↪  an effort to attend its presentation in a conference.
       3.5
       3 = Interesting: I might mention some points of this paper
       ↪  to others and/or attend its presentation in a
       ↪  conference if there's time.
       2.5
     2 = Potentially Interesting: this paper does not resonate with
       ↪  me, but it might with others in the *ACL community.
       1.5
       1 = Not Exciting: this paper does not resonate with me,
       ↪  and I don't think it would with others in the *ACL
       ↪  community (e.g. it is in no way related to
       ↪  computational processing of language).
       Overall Assessment
       If this paper was committed to an *ACL conference, do you
       ↪  believe it should be accepted? If you recommend
       ↪  conference, Findings and or even award consideration,
       ↪  you can still suggest minor revisions (e.g. typos,
       ↪  non-core missing refs, etc.).

       Outstanding papers should be either fascinating,
       ↪  controversial, surprising, impressive, or potentially
       ↪  field-changing. Awards will be decided based on the
       ↪  camera-ready version of the paper.
       We define "Best" as work that is particularly fascinating,
       ↪  controversial, surprising, impressive, and/or
       ↪  potentially field-changing.

       Main vs Findings papers: the main criteria for Findings
       ↪  are soundness and reproducibility. Conference
       ↪  recommendations may also consider novelty, impact and
       ↪  other factors.

       5 = Consider for Award: I think this paper could be
       ↪  considered for an outstanding paper award at an *ACL
       ↪  conference (up to top 2.5% papers).
       4.5 = Borderline Award
       4 = Conference: I think this paper could be accepted to an
       ↪  *ACL conference.
       3.5 = Borderline Conference
       3 = Findings: I think this paper could be accepted to the
       ↪  Findings of the ACL.
       2.5 = Borderline Findings
     2 = Resubmit next cycle: I think this paper needs substantial
       ↪  revisions that can be completed by the next ARR cycle.
       1.5 = Resubmit after next cycle: I think this paper needs
       ↪  substantial revisions that cannot be completed by the
       ↪  next ARR cycle.
     1 = Do not resubmit: This paper has to be fully redone, or it
       ↪  is not relevant to the *ACL community (e.g. it is in no
       ↪  way related to computational processing of language).
       Best paper justification
       If your overall assessment for this paper is either
       ↪  'Consider for award' or 'Borderline award', please
       ↪  briefly describe why.

  2. `overall/soundness_score` (float): Just the overall/soundness
    ↪  score as described in the ACL guidelines as a float.

  All interactions will be structured in the following way, with the
    ↪  appropriate values filled in.

  [[ ## paper ## ]]
  {paper}

  [[ ## review ## ]]
  {review}

  [[ ## overall/soundness_score ## ]]
  {overall/soundness_score}       # note: the value you produce
    ↪  must be a single float value

  [[ ## completed ## ]]

  In adhering to this structure, your objective is:
        Given an computer science research paper, generate a
       ↪  review of the paper
        and a numerical score approximating what you believe a
       ↪  peer reviewer would give the paper. Do not sugarcoat
       ↪  the review, honestly assess the proposed solution.
```

F.2 ICLR Prompt for Final Score Generation:

```
Your input fields are:
    1. `paper` (str): Computer science paper to review

Your output fields are:
    1. `review` (str):
        Reviewing a submission: step-by-step
        Summarized in one sentence, a review aims to determine whether a
       ↪  submission will bring sufficient value to the community and
       ↪  contribute new knowledge. The process can be broken down into
       ↪  the following main reviewer tasks:

        Read the paper: It's important to carefully read through the
       ↪  entire paper, and to look up any related work and citations
       ↪  that will help you comprehensively evaluate it. Be sure to
       ↪  give yourself sufficient time for this step.
        While reading, consider the following:
        Objective of the work: What is the goal of the paper? Is it to
       ↪  better address a known application or problem, draw attention
       ↪  to a new application or problem, or to introduce and/or
       ↪  explain a new theoretical finding? A combination of these?
       ↪  Different objectives will require different considerations as
       ↪  to potential value and impact.
        Strong points: is the submission clear, technically correct,
       ↪  experimentally rigorous, reproducible, does it present novel
       ↪  findings (e.g. theoretically, algorithmically, etc.)?
        Weak points: is it weak in any of the aspects listed in b.?
        Be mindful of potential biases and try to be open-minded about the
       ↪  value and interest a paper can hold for the entire ICLR
       ↪  community, even if it may not be very interesting for you.
        Answer four key questions for yourself, to make a recommendation
       ↪  to Accept or Reject:
        What is the specific question and/or problem tackled by the paper?
        Is the approach well motivated, including being well-placed in the
       ↪  literature?
        Does the paper support the claims? This includes determining if
       ↪  results, whether theoretical or empirical, are correct and if
       ↪  they are scientifically rigorous.
        What is the significance of the work? Does it contribute new
       ↪  knowledge and sufficient value to the community? Note, this
       ↪  does not necessarily require state-of-the-art results.
       ↪  Submissions bring value to the ICLR community when they
       ↪  convincingly demonstrate new, relevant, impactful knowledge
       ↪  (incl., empirical, theoretical, for practitioners, etc).
        Write and submit your initial review, organizing it as follows:
        Summarize what the paper claims to contribute. Be positive and
       ↪  constructive.
        List strong and weak points of the paper. Be as comprehensive as
       ↪  possible.
        Clearly state your initial recommendation (accept or reject) with
       ↪  one or two key reasons for this choice.
        Provide supporting arguments for your recommendation.
        Ask questions you would like answered by the authors to help you
       ↪  clarify your understanding of the paper and provide the
       ↪  additional evidence you need to be confident in your assessment.
        Provide additional feedback with the aim to improve the paper.
       ↪  Make it clear that these points are here to help, and not
       ↪  necessarily part of your decision assessment.
        Complete the CoE report: ICLR has adopted the following Code of
       ↪  Ethics (CoE). When submitting your review, you'll be asked to
       ↪  complete a CoE report for the paper. The report is a simple
       ↪  form with two questions. The first asks whether there is a
       ↪  potential violation of the CoE. The second is relevant only
       ↪  if there is a potential violation and asks the reviewer to
       ↪  explain why there may be a potential violation. In order to
       ↪  answer these questions, it is therefore important that you
       ↪  read the CoE before starting your reviews.

       Engage in discussion: The discussion phase at ICLR is different from
       ↪  most conferences in the AI/ML community. During this phase,
       ↪  reviewers, authors and area chairs engage in asynchronous
       ↪  discussion and authors are allowed to revise their
       ↪  submissions to address concerns that arise. It is crucial
       ↪  that you are actively engaged during this phase. Maintain a
       ↪  spirit of openness to changing your initial recommendation
       ↪  (either to a more positive or more negative) rating.
        Borderline paper meeting: Similarly to last year, the ACs are
       ↪  encouraged to (virtually) meet and discuss with reviewers
       ↪  only for borderline cases. ACs will reach out to schedule
       ↪  this meeting. This is to ensure active discussions among
       ↪  reviewers, and well-thought-out decisions. ACs will schedule
       ↪  the meeting and facilitate the discussion. For a productive
       ↪  discussion, it is important to familiarize yourself with
       ↪  other reviewers' feedback prior to the meeting. Please note
       ↪  that we will be leveraging information for reviewers who
       ↪  failed to attend this meeting (excluding emergencies).
        Provide final recommendation: Update your review, taking into
       ↪  account the new information collected during the discussion
       ↪  phase, and any revisions to the submission. (Note that
       ↪  reviewers can change their reviews after the author response
       ↪  period.)  State your reasoning and what did/didn't change
       ↪  your recommendation throughout the discussion phase.

    2. `overall_score` (float): Just the overall score as described in
      ↪  the ICLR guidelines as a float.
        10: Strong Accept: Often indicates the paper should be highlighted
       ↪  at the conference (e.g., oral presentation). Represents truly
       ↪  groundbreaking work or an excellent, top-tier paper.
```

```
        8: Accept: Represents a good, solid paper that clearly meets the
       ↪  acceptance criteria.
       6: Weak Accept / Marginally Above Threshold: Indicates the paper is
       ↪  likely acceptable, but perhaps less impactful or polished than
       ↪  higher-rated papers. The reviewer leans towards acceptance.
        5: Weak Reject / Marginally Below Threshold: Indicates the paper
       ↪  has merits but falls slightly short of the acceptance bar.
       ↪  The reviewer leans towards rejection but might be swayed
       ↪  during discussion.
        3: Reject: Indicates the paper is not considered good enough for
       ↪  acceptance due to significant flaws, lack of novelty, or
       ↪  other issues.
        1: Strong Reject: Indicates the paper has major flaws, is clearly
       ↪  unsuitable for the conference, or perhaps should not have
       ↪  been submitted in its current state.

    All interactions will be structured in the following way, with the
    ↪  appropriate values filled in.

    [[ ## paper ## ]]
    {paper}

    [[ ## review ## ]]
    {review}

    [[ ## overall_score ## ]]
    {overall_score}        # note: the value you produce must be a single
    ↪  float value

    [[ ## completed ## ]]

    In adhering to this structure, your objective is:
            Given an computer science research paper, generate a review of
            ↪  the paper
             and a numerical score approximating what you believe a peer
            ↪  reviewer would give the paper. Do not sugarcoat the
            ↪  review, honestly assess the proposed solution.
```

F.3 GoodReads Prompt for Final Score Generation:

```
Your input fields are:
        1. `book_summary` (str): Book summary to review

Your output fields are:
        1. `review` (str):
            Review Form
                Book Summary
               Provide a brief summary of the book's plot and main themes.

                Strengths
                What are the major strengths of this book? Consider
                ↪  elements like:
                - Writing style and prose
                - Character development
                - Plot structure and pacing
                - World-building (for fiction)
                - Research and accuracy (for non-fiction)
                - Originality and creativity
                - Emotional impact
                - Themes and messages

                Weaknesses
                What aspects of the book could be improved? Consider:
                - Plot holes or inconsistencies
                - Weak character development
                - Pacing issues
                - Writing style problems
                - Research gaps (for non-fiction)
                - Unoriginal elements
                - Unresolved plot threads
                - Unclear themes or messages

                Recommendation
                Who would enjoy this book? What type of reader would find
                ↪  it most appealing?

        2. `overall_score` (float): Rate the book on a scale of 1-5 stars,
        ↪  where:
            5.0 = It was amazing
            4.0 = Really liked it
            3.0 = Liked it
            2.0 = It was ok
            1.0 = Did not like it

        All interactions will be structured in the following way, with the
        ↪  appropriate values filled in.

        [[ ## book_summary ## ]]
        {book_summary}

        [[ ## review ## ]]
        {review}

        [[ ## overall_score ## ]]
        {overall_score}        # note: the value you produce must be a
        ↪  single float value

        [[ ## completed ## ]]

    In adhering to this structure, your objective is:
            Given a book, generate a review and rating that reflects
            ↪  your honest assessment of its quality.
```