MemoNoveltyAgent: A Historical Research Memory-Aware Agent Workflow for Paper Novelty Assessment

Jiajun Hou[5*] **Hexuan Deng**[1,3*] **Wenxiang Jiao**[2]
Xuebo Liu[1†] **Xiaopeng Ke**[1] **Derek F. Wong**[4] **Min Zhang**[1]

[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
[2]Xiaohongshu Inc. [3]Zhongguancun Academy, Beijing, China
[4]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau, China
[5]The Chinese University of Hong Kong, Shenzhen, China
{jiajunhou738,hxuandeng,wenxiangjiaonju,xiaopk7}@gmail.com
{liuxuebo,zhangmin2021}@hit.edu.cn
derekfw@um.edu.mo

arXiv:2603.20884v4 [cs.CL] 30 Aug 2026

Abstract

To alleviate the heavy burden of paper screening, researchers increasingly rely on existing AI agents, such as AI reviewers or DeepResearch, for paper evaluation and novelty assessment. However, lacking specialized mechanisms for processing scholarly literature, their analyses often produce superficial results with noticeable deficiencies in quality. To bridge this gap, we introduce MemoNoveltyAgent, a multi-agent system designed to generate comprehensive and faithful novelty reports. Beyond retrieving concrete prior-paper evidence via RAG, our system incorporates a high-level abstract memory constructed from large-scale scholarly corpora. This memory organizes research into hierarchical trees to distill field-specific evolutionary trajectories, thereby providing a broader historical context. Furthermore, we decompose papers into discrete novelty points for fine-grained analysis and retrieval, while employing a self-validation mechanism to improve report faithfulness. Finally, to address the evaluation challenges of such open-ended generation tasks, we propose a RAG-augmented checklist evaluation method that enables reliable and evidence-grounded assessments. Extensive experiments demonstrate that MemoNoveltyAgent outperforms GPT-5 DeepResearch by 13.69%. Code and demo are available at https://github.com/SStan1/MemoNoveltyAgent.

1 Introduction

Manually verifying the novelty of a paper is highly time-consuming, as it requires tracing related work and judging whether each claimed contribution is truly new. With the rapid proliferation of scientific literature, this manual overhead has increased drastically, making high-quality peer review increasingly unsustainable. Therefore, reliable automated novelty assessment has become indispensable, serving both formal peer reviewers and scholars who need to efficiently evaluate new papers.

Existing automated assistance mainly falls into two categories. General-purpose research agents, such as DeepResearch, are often repurposed for paper novelty evaluation despite being designed for general web-retrieval tasks. On the other hand, automated AI reviewers can also be employed for this assessment, though they treat novelty analysis merely as a sub-task within the broader peer-review process. However, because both approaches lack robust and sophisticated mechanisms for retrieving and processing scholarly literature, their generated analysis reports often exhibit noticeable deficiencies, particularly in terms of completeness, depth, and faithfulness, rendering the content generic, superficial, and susceptible to factual hallucinations.

These limitations arise from several primary factors. First, by processing a manuscript as an undivided whole, current workflows often obscure specific novelty points and overlook key contributions. Second, during evidence gathering, standard semantic retrieval frequently fails to retrieve earlier or implicit works that use different terminology. Third, when evaluating the gathered context, these systems rely entirely on the LLM's internal reasoning without an objective baseline. Unlike human experts who use historical literature as a reference to evaluate new submissions, LLMs lack such benchmarks, which hinders in-depth analysis and leads to generic evaluations. Finally, without explicit validation mechanisms, these systems are susceptible to hallucinations, thereby compromis-

*Equal contribution.
†Corresponding author.

ing the reliability of the generated analysis.

To address these limitations, we introduce MemoNoveltyAgent, a multi-agent system designed to generate comprehensive and faithful novelty reports. Specifically, MemoNoveltyAgent constructs a large-scale pre-computed abstract memory from a massive literature database. Through clustering and recursive refinement, this memory organizes ICLR research into hierarchical trees that distill the developmental histories of fine-grained subfields, thereby providing broader historical context together with expert-reviewer evaluation results. In addition, MemoNoveltyAgent builds a contextual literature database from the submission's citation network as an auxiliary reference. During evaluation, the workflow decomposes a target submission into discrete novelty points and retrieves relevant evidence from both the contextual literature database and the abstract memory. The retrieved evidence then guides report generation, followed by an independent self-validation protocol to reduce factual hallucinations and improve citation integrity. Finally, to robustly evaluate open-ended generation, we propose a checklist-based evaluation method featuring dense question coverage and a RAG-integrated verification mechanism. Extensive experiments demonstrate that MemoNoveltyAgent achieves state-of-the-art performance, outperforming the strong general-purpose baseline GPT-5 DeepResearch by 13.69% on the main benchmark, and by 10.35% on a broader cross-venue dataset. Our main contributions are:

- We propose MemoNoveltyAgent, a multi-agent workflow for comprehensive and faithful paper novelty report generation.
- We build a historical research memory and combine it with citation-network-based retrieval, point-wise decomposition, and self-validation to improve report faithfulness.
- We introduce a RAG-integrated checklist evaluation method and validate its effectiveness and reliability through human alignment.

2 Related Work

DeepResearch. To overcome the inherent limitations of static parametric memory (Mallen et al., 2022; Deng et al., 2025), modern LLMs such as ChatGPT (OpenAI, 2023), Gemini (Team, 2023), and Kimi (Moonshot AI, 2024) have integrated tools like web browsing capabilities, allowing models to access external knowledge bases, thereby grounding their generation in up-to-date information (Lewis et al., 2020; Nakano et al., 2021; Deng et al., 2023; Chen et al., 2025).

To further address complex inquiries, "DeepResearch" modes have been proposed (Google, 2024; OpenAI, 2025; xAI, 2025; Perplexity, 2025). Based on observations of their generated outputs and opinions from recent scholars (Zhang et al., 2025a; Shi et al., 2025), the core components of this paradigm generally include: first, *planning*, which decomposes complex queries into sub-questions and constructs research plans (Gu et al., 2025); second, *web exploration*, where agents search for relevant information online, accompanied by continuous iteration and reflection (Nakano et al., 2021; He et al., 2024; Ke et al., 2025; Deng et al., 2026a; Rao et al., 2025); and finally, *report generation*, which synthesizes the collected evidence into coherent responses (Li et al., 2025b).

While designed for general queries and applicable to novelty assessment, their web-centric retrieval struggles with scholarly literature. Specifically, they are easily distracted by non-academic content like blogs, and their limited search breadth fails to capture the comprehensive research landscape.

AI Reviewer. Automated peer review systems aim to address the complex nature of the reviewing process (Zhuang et al., 2025). To capture interactive dynamics, recent agent-based frameworks simulate multi-turn dialogues (Jin et al., 2024), treat reviewing as a long-context task (Tan et al., 2024), or utilize multi-agent collaboration to enhance analytical depth (Jin et al., 2024; Tian et al., 2025; Wang et al., 2026). Furthermore, reinforcement learning aligns automated reviewers with human preferences through iterative loops (Weng et al., 2025) or scoring optimization (Zeng et al., 2025), while knowledge-enhanced models provide explainable critiques (Wang et al., 2020). Finally, multimodal systems integrate visual encoders to interpret figures and tables for holistic assessment (Hong et al., 2025), supported by benchmarks (Gao et al., 2025; Roberts et al., 2024; Deng et al., 2026b).

Recent studies have begun to address scientific novelty assessment more directly. Shahid et al. (2025) use literature retrieval and reranking to assess the novelty of scientific ideas. da Silva et al. (2025) combine structured paper representations

with citation-based and semantic retrieval for interactive novelty assessment.

Despite these advances, existing systems still face certain limitations. First, directly processing the entire paper heavily increases complexity. Second, relying on LLMs to make direct judgments from raw information without references often leads to generic analyses that overlook critical details.

3 MemoNoveltyAgent Workflow

We now detail MemoNoveltyAgent, an automated agent workflow that takes a paper title as input, downloads related content, and conducts novelty analysis. As shown in Figure 1, it consists of four main stages: Literature Database Construction, Historical Research Memory Construction, Point-Wise Report Generation, and Faithfulness-Enhanced Self-Validation.

3.1 Literature Database Construction

Given a target paper P, MemoNoveltyAgent constructs a local literature database scoped within its citation neighborhood. We begin by collecting all first-order references (i.e., the papers directly cited by P). We then expand this network to include second-order references (i.e., the papers cited by the first-order references). Because the raw second-order set can be noisy and excessively large, we prioritize papers based on structural relevance: papers cited by more first-order references are ranked higher, and more recent papers are preferred when the co-occurrence signal is tied. Through this process, each target paper is associated with a capped database of prioritized full-text PDFs.

To facilitate precise information retrieval from this constructed database, MemoNoveltyAgent implements a hybrid Retrieval-Augmented Generation (RAG) pipeline. The detailed implementation of this RAG pipeline is provided in Appendix A.4.

3.2 Historical Research Memory Construction

While the literature database primarily focuses on the most relevant and granular literature details, we construct an offline Historical Research Memory to provide a broader and deeper historical context. It systematically stores historical novelty points, development trajectories, and peer review comments to offer an objective reference benchmark. This module serves both as an internal knowledge anchor providing information to the LLM agents, and as a standalone, interpretable tool for researchers to directly trace the chronological development of specific fields.

Historical Novelty Extraction. We collect accepted ICLR papers from 2017 to 2025 and apply the same novelty-point extraction method used in the main workflow. This yields approximately 32K historical novelty points from more than 8K accepted papers. Each extracted point is stored alongside its source paper title, year, summary, and available OpenReview reviewer comments.

Coarse-to-Fine Theme Tree Construction. To avoid the high computational costs of relying solely on LLMs to categorize massive novelty points, we adopt a two-stage clustering strategy that balances cost and accuracy. First, we embed all novelty points and use DBSCAN to form roughly 400 coarse clusters, averaging about 80 points per cluster. Since many coarse clusters still contain multiple sub-directions, we then employ an LLM for fine-grained secondary classification. The LLM recursively refines any cluster with more than 15 items, splitting them into more granular child themes until each leaf node contains at most 15 closely related points, ultimately establishing a well-structured theme tree. A theme tree example is shown in Appendix A.2.

Writing Leaf and Aggregated Memories. We conceptualize each leaf node as a distinct micro-domain formed by closely related innovation points. For each micro-domain, the LLM synthesizes the innovation texts, summaries, and reviewer feedback to construct a detailed evolutionary trajectory. This record comprises two layers: (1) a specific content and novelty analysis for each individual innovation point, and (2) a chronological summary of the micro-domain's overall evolution. For parent nodes, the LLM aggregates the summaries of their child nodes to generate a broader evolutionary trajectory. Figure 2 provides a concrete example of a parent node's evolutionary trajectory.

3.3 Point-Wise Report Generation

To ensure comprehensive evaluation and structural alignment with the Historical Research Memory, we adopt a point-by-point strategy. A dedicated agent first decomposes the manuscript into individual novelty points, which subsequent agents then analyze independently.

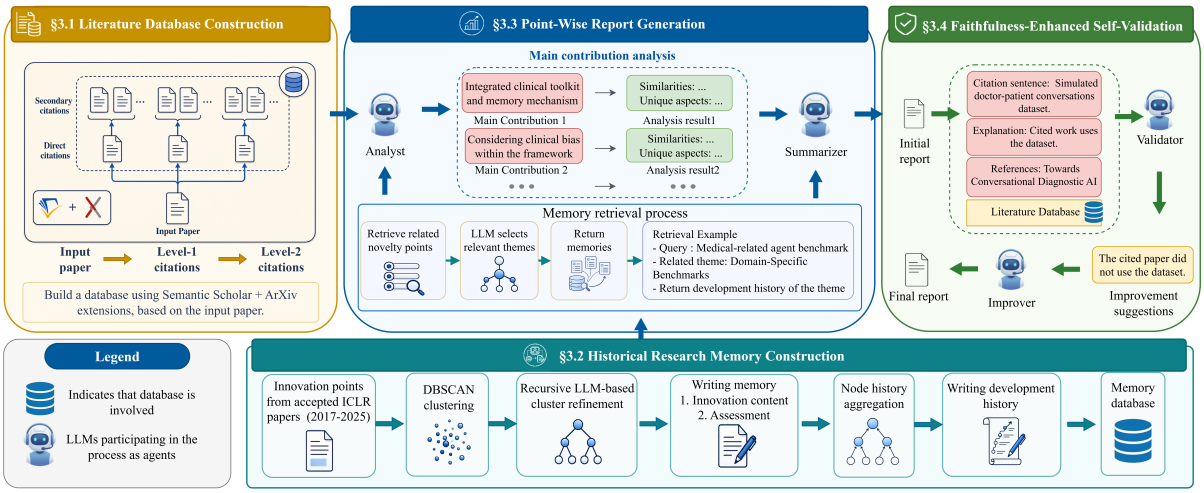


Figure 1: The MemoNoveltyAgent workflow. The system consists of four primary modules: (§3.1) constructing a citation-based literature database; (§3.2) building an offline historical research memory database from ICLR accepted papers; (§3.3) generating a point-wise report by combining local RAG and memory retrieval; and (§3.4) performing faithfulness-enhanced self-validation to output the final report.

Splitting Agent. It first extracts a set of discrete, core novelty points $\{i_1, ..., i_N\}$ from the target paper P. For each extracted novelty point i_k, the agent generates multiple targeted queries and retrieves supporting evidence from the literature database through the RAG pipeline.

Analyst Agent. Using the extracted novelty points i_k and their retrieved contexts $\mathcal{D}_k$, the Analyst Agent conducts a rigorous comparative analysis. First, the target novelty point is embedded to retrieve the closest historical points from the Historical Research Memory, mapping them to their corresponding leaf themes. The LLM then selects the one or two most relevant themes, supplying the Analyst Agent with the specific innovation details and local analyses of all historical points within these nodes. By leveraging this macro-level organization of development trajectories, we overcome the limitations of standard semantic retrieval. This approach successfully recalls relevant papers that might otherwise be missed due to terminology variations or implicit phrasing, ensuring a comprehensive context for technical comparison.

Summarizer Agent. The Summarizer Agent synthesizes the final output and drafts the **Novelty Summary**. To distill core insights, the agent retrieves relevant context from the Historical Research Memory, specifically the novelty assessment for each innovation point, the local development history of the matched leaf theme, and the broader developmental history of its parent node. Because these development trajectories inherently incorporate historical expert peer reviews, the Summarizer Agent assimilates these critical perspectives to better simulate human expert evaluation standards during macro-summarization. Finally, it aggregates a baseline factual overview of the paper, the Point-Wise Novelty Analysis, and the Novelty Summary to construct the complete **Initial Report**.

3.4 Faithfulness-Enhanced Self-Validation

To ensure report credibility and mitigate hallucinations, MemoNoveltyAgent incorporates a self-validation module that verifies and corrects the generated content against the source materials.

Validator and Improver Agents. To prevent source distortion caused by the lack of independent verification, the **Validator Agent** systematically constructs **citation pairs**, linking report statements directly to their full-text sources. It scrutinizes these pairs for strict faithfulness and generates actionable feedback for any identified misinterpretations. Guided by this feedback, the **Improver Agent** performs directional corrections to rectify hallucinations. Following these targeted edits, it conducts a comprehensive polishing pass to enhance linguistic fluency, terminological consistency, and logical flow, ultimately guaranteeing the rigorous faithfulness of the final analysis.

Output Report. The final output report R follows a structured, three-section format. First, the Paper Content Summary provides a concise fac-

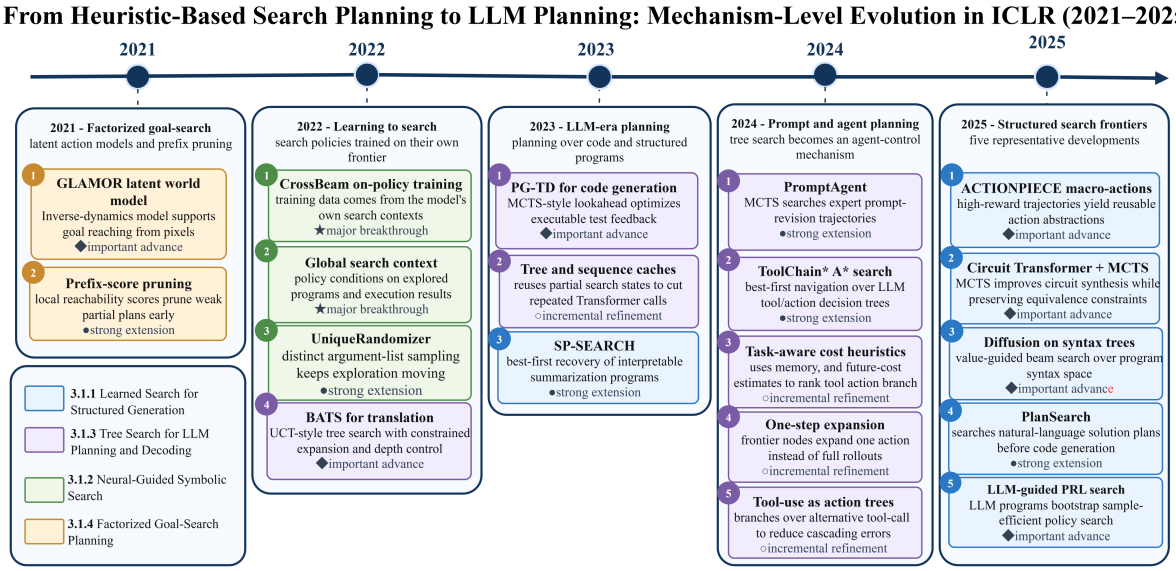


Figure 2: A concrete example of the evolutionary trajectory of a parent node.

tual overview of P, capturing its core objectives and key technical approaches. Second, the Point-wise Novelty Analysis conducts a comparative analysis against retrieved prior work for each extracted novelty point, identifying technical similarities and highlighting the genuine innovations. Finally, the Novelty Summary synthesizes these findings into a dialectical assessment that carefully distinguishes genuine breakthroughs from incremental modifications. Specific cases are showcased in Appendix A.8.

3.5 Checklist-based Evaluation

Evaluating open-ended novelty reports is challenging: manual assessment is unscalable, and standard metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) fail to capture semantic depth and factual accuracy. Even recent LLM-as-a-judge approaches struggle to provide reliable evaluations for such nuanced tasks (Li et al., 2025a; Wang et al., 2024). To address this, we develop an automated, **RAG-enhanced checklist-based evaluation method**. By decomposing holistic assessment into specific, verifiable criteria (Yes/No questions), this approach forces the model to make judgments based on atomic, concrete facts, thereby minimizing subjectivity and enhancing reliability (Cook et al., 2024).

Checklist Construction. Drawing inspiration from CheckEval (Lee et al., 2025), we construct an evaluation checklist spanning five key dimensions: Fluency assesses linguistic quality and clarity of expression; Faithfulness evaluates factual alignment with source materials to prevent hallucinations; Completeness measures whether the report omits any essential information that should be considered; Effectiveness confirms strict adherence to the predefined template and focus on core analysis; and Depth evaluates intellectual rigor and the inclusion of specific technical details. We construct this checklist, comprising a total of 118 questions, through LLM generation followed by human verification. The detailed construction process is provided in Appendix A.10. We then compute, for each model, the proportion of checklist questions answered with yes and rescale this proportion to a 0–10 range as the model's final score.

RAG Integration. For faithfulness and other criteria requiring external evidence, we connect the evaluator to the database constructed in §3.1, enabling the LLM to retrieve relevant paper content during verification. Detailed prompts and implementation specifics are provided in Appendices A.4 and A.10. Furthermore, by integrating a RAG system, this evaluation framework can be extended to a wider range of scenarios, providing a viable paradigm for assessing open-ended text generation in the future.

4 Experiments

To evaluate the performance of our method, we design our experiments to answer the following research questions: **RQ1**: How does MemoNoveltyAgent perform compared to existing automated

Models	Score (0-10) ↑					
	Completeness	Depth	Effectiveness	Faithfulness	Fluency	Overall
Web Research						
Kimi-2	7.57	7.01	9.22	4.85	9.77	7.68
GPT-5 Thinking	7.92	7.21	9.67	4.90	9.87	7.91
Gemini-2.5-Flash DeepResearch	8.18	8.03	9.58	4.00	8.61	7.68
GPT-5 DeepResearch	8.06	7.72	9.45	6.06	9.59	8.18
AI Reviewer						
DeepReview	8.39	8.10	9.51	7.56	9.85	8.68
AgentReviewer	8.55	8.07	9.28	6.46	**9.91**	8.45
MemoNoveltyAgent	**9.20**	**8.80**	**9.93**	**8.76**	9.83	**9.30**

Table 1: Main evaluation results on the 50-paper ICLR dataset (0–10 scale). Best results are highlighted.

Models	Proxy $(n-1)$		Proxy (n)	
	MSE↓	MAE↓	MSE↓	MAE↓
Gemini-2.5-Flash-Nothinking	**0.28**	**0.44**	**0.19**	**0.37**
Gemini-2.5-Flash	0.37	0.49	0.26	0.41
GPT-4o-Mini	0.72	0.79	0.50	0.66
Claude-3-Haiku-20240307	0.93	0.77	0.64	0.64
DeepSeek-V3-250324	1.11	0.99	0.77	0.82
O4-Mini-2025-04-16	1.46	1.06	1.01	0.88

Table 2: LLM Consistency result on a 32-paper subset.

tools? **RQ2**: How reliable and effective is the proposed automated evaluation method? **RQ3**: What is the individual contribution of each core component, and how well does the framework generalize across different scenarios?

4.1 Experimental Setup

Dataset Construction. We constructed an evaluation dataset based on ICLR 2025 submissions from OpenReview. To ensure diversity and representativeness, we select 50 target papers via stratified sampling based on a weighted combination of reviewer overall scores and originality scores. These target papers are evenly distributed across four specific score ranges (0-3, 3-5, 5-7, and 7-10) to ensure a balanced evaluation. Furthermore, the selection maximizes the coverage of research fields, encompassing papers from 20 distinct research topics. For each target paper, we constructed a literature database comprising 200 PDF files.

Baseline Models. We compare MemoNoveltyAgent against two categories of baselines: *Web Research* and *AI Reviewer*. The web research systems include GPT-5 Thinking, Kimi-2 (Moonshot AI, 2024), GPT-5 DeepResearch (OpenAI, 2025), and Gemini-2.5-Flash DeepResearch (Google, 2024). These systems retrieve online information to generate novelty reports, with some of them supporting more advanced deep research capabilities such as multi-step web exploration. For all web research systems, we conduct testing through the services provided on their official websites. The AI reviewer systems include AgentReviewer (Jin et al., 2024) and DeepReview (Zhu et al., 2025), which use agent-based paper reviewing frameworks and proprietary academic databases to generate novelty reports.

4.2 Main Results

Table 1 presents the main evaluation results across the five defined dimensions on our primary testbed. MemoNoveltyAgent achieves the highest overall score of 9.30, outperforming the strongest baseline, DeepReview, by a margin of 0.62 points.

Factual Accuracy and Reliability. MemoNoveltyAgent achieves the highest Faithfulness score of 8.76, demonstrating that our generated results are highly reliable with minimal hallucinations. In contrast, web-based baselines generally exhibit lower faithfulness as they are susceptible to data volatility and unverified claims.

Analytical Depth and Critical Review Perspectives. MemoNoveltyAgent achieves the highest Depth score of 8.80. This indicates that our system provides richer technical details and objective evaluations. Unlike baselines that often generate generic summaries from isolated text, our framework retrieves fine-grained technical features. Furthermore, by incorporating expert review perspectives, the system generates more professional and critical technical assessments.

Breadth of Literature Coverage. MemoNoveltyAgent secures a leading Completeness score of 9.20. This evaluation result underscores that our generated reports suffer from fewer omissions

Models	Score (1–5) ↑											
	Completeness		Depth		Effectiveness		Faithfulness		Fluency		Overall	
	H	AI	H	AI	H	AI	H	AI	H	AI	H	AI
DeepReview	3.875	4.17	3.50	4.16	4.75	4.78	3.75	4.04	**4.875**	**4.91**	4.15	4.41
GPT-5 DeepResearch	3.50	4.15	3.25	3.99	4.50	4.67	3.00	3.15	4.50	4.78	3.75	4.15
MemoNoveltyAgent	**4.25**	**4.53**	**4.50**	**4.34**	**4.88**	**4.89**	**4.25**	**4.32**	4.50	4.87	**4.48**	**4.59**

Table 3: Human (H) and AI evaluation result on a 32-paper subset. AI scores are scaled to the 1–5 human rubric.

and achieve broader literature coverage. Standard baselines frequently miss crucial prior works due to evolving terminologies across different eras. Our framework alleviates this bottleneck through a dual-source retrieval mechanism, which couples the local literature database with historical technological trajectories in our memory module, maximizing information recall.

Structural and Linguistic Proficiency. Regarding structural compliance and readability, the evaluated models generally maintain high performance standards, with most models scoring above 9.20 in both dimensions. This indicates that current large language models are inherently capable of satisfying baseline linguistic fluency and format adherence. Consequently, the critical performance differentiation among automated reviewing systems is driven primarily by factual correctness and analytical thoroughness rather than surface presentation.

4.3 Evaluation Method Validity

We aim to validate the effectiveness of our proposed evaluation method. Due to the absence of an absolute gold-standard answer, we employ two distinct validation approaches: human alignment and LLM consistency. Inspired by recent automated benchmarking methodologies, the LLM consistency approach utilizes multiple language models as evaluators, taking their average consensus score as a proxy gold standard. If our evaluation method demonstrates high alignment with human judgments and exhibits minimal deviation from the LLM consensus, it is considered robust and reliable.

LLM Consistency. To rigorously evaluate the consistency of our automated evaluation method, we conducted a study on a representative subset of 32 papers. Specifically, we collected the novelty reports generated by four baselines—GPT-5 Online, GPT-5 DeepResearch, MemoNoveltyAgent, and DeepReview—to serve as the evaluation targets. To mitigate potential systematic bias stemming from a single model architecture, we deliberately selected evaluator models from four distinct vendors, as illustrated in Table 2. To establish proxy ground truths (GT) for quantifying scoring consistency, we employed two distinct aggregation strategies: (1) the $n-1$ strategy, where the proxy GT for a specific evaluator model is defined as the average score assigned by the remaining models; and (2) the n strategy, where the proxy GT is calculated by averaging the scores from all participating models. We then computed the Mean Absolute Error (MAE) and Mean Squared Error (MSE) to quantify the mathematical alignment between each model's predicted score and the collective proxy GT.

High Consistency Across Models. Table 2 presents the comparative results of various evaluator models. Notably, Gemini-2.5-Flash-Nothinking emerges as the top-performing model, achieving a remarkably low MSE of **0.28** and an MAE of **0.44** under the $n-1$ strategy. This superior alignment indicates that the model closely adheres to the collective consensus. Driven by this exceptional stability, we adopt it as the primary evaluator for all subsequent formal assessments, which further substantiates that our overall evaluation method yields stable and reliable outcomes.

Human Evaluation Alignment. To verify if our automated proxy metrics anchor well with professional human standards, we sampled 8 papers from each score interval to build a 32-paper subset for manual grading. As detailed in Table 3, human experts and our automated checklist evaluation method exhibit remarkably tight alignment. From a pairwise comparison perspective, except for language fluency which results in a statistical tie, the preference rankings between the automated evaluator and human experts are completely consistent across all other core dimensions. Compared to the automated evaluator, human reviewers observe a wider margin in analytical depth and demonstrate higher acceptance for MemoNoveltyAgent.

Models	Score (0-10) ↑					
	Completeness	Depth	Effectiveness	Faithfulness	Fluency	Overall
MemoNoveltyAgent	**9.20**	**8.80**	**9.93**	**8.76**	9.83	**9.30**
w/o Memory	8.76	8.49	9.79	8.43	9.66	9.03
Direct Generation	8.35	7.44	8.29	7.78	**9.86**	8.34

Table 4: Ablation results on the main 50-paper ICLR dataset. "w/o Memory" only removes the memory module. "Direct Generation" removes memory, point-wise analysis, and self-validation.

Method	Completeness	Depth	Reviewer Helpfulness	Faithfulness
Kimi-2 Online	3.06	3.14	3.02	4.74
GPT-5 Thinking	3.42	3.42	3.44	4.88
Gemini-2.5-Flash DeepResearch	3.40	3.62	3.40	4.98
GPT-5 DeepResearch	3.66	3.82	3.68	4.98
AgentReviewer	4.00	4.04	3.98	**5.00**
DeepReview	4.00	4.22	4.02	4.96
MemoNoveltyAgent	**4.47**	**4.85**	**4.60**	4.98

Table 5: Reviewer-reference evaluation against an approximate ground truth constructed from novelty-related comments in the original ICLR 2025 OpenReview reviews. Scores are reported on a 1–5 scale, and the best result in each dimension is highlighted.

This strongly validates the effectiveness of our automated evaluation method and its high alignment with human cognition.

Reviewer-Reference Evaluation. Beyond comparing automated scores with human ratings, we directly evaluate report content against an external reference constructed from real reviewer comments. This experiment covers all 50 papers in the main ICLR 2025 dataset. For each paper, GPT-5.4 reads all original OpenReview reviews and extracts novelty-related comments as structured points, including contributions recognized as novel and aspects considered insufficiently novel or incremental. We treat these structured reviewer points as an approximate ground truth for the novelty issues that a useful report should address.

We then use GPT-5.4 to compare each complete generated report against these extracted reviewer comments. The presentation order of reports from different systems is randomized to reduce position bias. Each report is evaluated on a 1–5 scale along four dimensions: *Completeness*, measuring coverage of reviewer-identified novelty comments; *Depth*, measuring whether the report discusses the same fine-grained technical mechanisms; *Reviewer Helpfulness*, measuring how effectively the report reduces the need for manual literature retrieval and comparison; and *Faithfulness*, measuring whether the report contains explicit factual conflicts with the reviewer comments.

As shown in Table 5, MemoNoveltyAgent ranks first in Completeness, Depth, and Reviewer Helpfulness, achieving scores of 4.47, 4.85, and 4.60, respectively. These results show that its reports more fully cover reviewer-identified novelty issues, provide more technically detailed comparisons, and offer greater practical support for reviewers. Faithfulness scores are uniformly high across systems and differ only slightly, partly because original reviewer comments are typically concise and provide limited information from which explicit factual conflicts can be identified. Overall, this experiment compares the generated reports against novelty-related assessments provided by human reviewers as an approximate ground truth. The resulting rankings are consistent with those of the preceding evaluations, thereby complementing the existing validation analyses and making the overall evaluation framework more comprehensive.

4.4 Ablation Study and Component Analysis

To assess the distinct performance contribution of each component, we conducted incremental ablation experiments on our main dataset. Table 4 details the quantitative variations.

Impact of Base Framework Architecture. Compared to "Direct Generation", adopting the base agentic framework (w/o Memory) increases the overall score from 8.34 to 9.03, reflecting the structural advantages of our workflow. This base framework improves performance primarily through two core mechanisms. First, the notable

Models	Score (0–10) ↑					
	Completeness	Depth	Effectiveness	Faithfulness	Fluency	Overall
GPT-5.4 Online	6.30	3.87	9.49	5.18	**9.73**	6.91
GPT-5.4 DeepResearch	7.64	7.42	9.66	6.91	9.40	8.21
DeepReview	8.42	7.72	9.79	7.30	9.54	8.55
MemoNoveltyAgent	**8.82**	**8.50**	**9.94**	**8.57**	9.49	**9.06**

Table 6: Evaluation result on the cross-venue dataset, including 50 papers from ACL, CVPR, ACM MM, TPAMI, and JMLR.

improvement in Faithfulness (+0.65) benefits from the closed-loop self-validation mechanism, which effectively filters hallucinations. Second, the increases in Depth (+1.05) and Completeness (+0.41) are driven by the point-wise analysis paradigm. By decomposing the overall report generation process into independent, fine-grained novelty points, the system retrieves richer context to provide specific technical details and ensure broader coverage of relevant information.

Impact of Historical Research Memory. Integrating the historical research memory pushes the overall rating from 9.03 to its peak of 9.30. The empirical benefits of this module are reflected in two dimensions. For literature Completeness (+0.44), the memory acts as a vital supplementary source alongside the literature database, establishing a robust dual-source retrieval mechanism that ensures broader information recall. For analytical Depth (+0.31), the incorporation of historical peer-review feedback equips the framework with an expert reference benchmark. Rather than relying solely on raw paper texts, this professional baseline calibrates the language model's cognitive scale, thereby significantly enhancing its independent, dialectical reasoning and critical thinking capabilities.

4.5 Cross-Venue Generalization Analysis

To evaluate the performance of MemoNoveltyAgent across different research domains and writing styles, we establish a cross-venue dataset.

Cross-Venue Dataset Setup. We construct a cross-venue dataset consisting of 50 papers sampled evenly from five venues: ACL, CVPR, ACM MM, TPAMI, and JMLR. It covers natural language processing, computer vision, multimodal learning, and statistical machine learning, spanning both conference and journal publication formats.

Performance Stability. Table 6 shows the cross-venue evaluation results. MemoNoveltyAgent achieves the highest overall score of 9.06, outperforming DeepReview by 0.51 points, and maintains a consistent lead across completeness (8.82), depth (8.50) and faithfulness (8.57). While absolute scores show a slight overall decline compared to the native ICLR environment—primarily because our historical memory is constructed using previously accepted ICLR papers, which do not strictly align with the distributions of papers from these other venues, resulting in cross-venue domain shifts—the framework maintains reliable reporting performance without specialized fine-tuning.

5 Conclusion

In this work, we present MemoNoveltyAgent, a multi-agent workflow for analyzing the novelty of academic papers and generating comprehensive, faithful reports. At its core, the system introduces a historical research memory assisted by a point-wise analysis and self-validation architecture. To support a reliable evaluation of this open task, we also propose a checklist-based evaluation method. We hope MemoNoveltyAgent can serve as a reliable tool for high-quality novelty analysis, reducing the cost of paper screening and helping researchers quickly identify truly original papers.

Limitations

We acknowledge a few limitations in our current approach. First, our proposed agent framework relies heavily on open-access academic papers, which is why our current evaluation and application are primarily focused on the Computer Science (CS) domain. In several other disciplines, such as biological sciences, a substantial portion of academic literature is behind paywalls and lacks public retrieval methods. This makes it exceedingly difficult to acquire a sufficient volume of full-text papers required to construct the historical research memory and the local literature database. To extend our framework to these restricted domains, future work will need to explore alternative techni-

cal routes (e.g., partnering with academic publishers or developing lightweight verification strategies based solely on abstracts and metadata). Second, for every target input, our framework dynamically constructs a local literature database, which necessitates downloading a large volume of PDF files in real time. This process imposes a significant demand on network resources. Consequently, on servers or environments with poor network connectivity, the PDF acquisition phase can consume a considerable amount of time, thereby bottlenecking the overall efficiency. Finally, the execution time for processing a single paper is relatively long. The complex multi-agent workflow requires each agent to occupy a substantial amount of processing time for extensive retrieval, reading, and reasoning. However, this limitation is primarily evident in sequential, single-instance runs; when processing large batches of papers, the overall time cost can be effectively mitigated through parallel execution and distributed deployment.

Acknowledgments

This work is supported in part by the Zhongguancun Academy (Grant No.s C20250203), the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

Ethics Statement

Our work adheres to the ACL Ethics Policy and uses only publicly available scholarly resources and datasets for reproducibility. Since the data come from open-access academic resources, we do not collect private data from individuals. For the human annotation involved in this work, participants were informed of the purpose of the study and how their annotations would be used before participation. LLMs may exhibit racial and gender biases, so we strongly recommend users assess potential biases before applying the system in specific contexts. Additionally, due to the difficulty of controlling LLM outputs, users should be cautious about hallucinations and unreliable generations.

References

Guanhua Chen, Yutong Yao, Derek F. Wong, Xuebo Liu, and Lidia S. Chao. 2025. SGIC: A self-guided iterative calibration framework for RAG. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.

Jonathan Cook, Tim Rocktäschel, Jakob N. Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve LLM evaluation and generation. CoRR, abs/2410.03608.

Italo Luis da Silva, Hanqi Yan, Lin Gui, and Yulan He. 2025. Graphmind: Interactive novelty assessment system for accelerating scientific discovery. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - System Demonstrations, Suzhou, China, November 4-9, 2025, pages 286–294. Association for Computational Linguistics.

Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. Improving simultaneous machine translation with monolingual data. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 12728–12736. AAAI Press.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Jing Li, Min Zhang, and Zhaopeng Tu. 2025. DRPruning: Efficient large language model pruning through distributionally robust optimization. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29152–29173, Vienna, Austria. Association for Computational Linguistics.

Hexuan Deng, Wenxiang Jiao, Xuebo Liu, Jun Rao, and Min Zhang. 2026a. REA-RL: Reflection-aware online reinforcement learning for efficient reasoning. In The Fourteenth International Conference on Learning Representations.

Hexuan Deng, Xiaopeng Ke, Yichen Li, Ruina Hu, Dehao Huang, Derek F. Wong, Yue Wang, Xuebo Liu, and Min Zhang. 2026b. Cocoreviewbench: A completeness- and correctness-oriented benchmark for AI reviewers. In Forty-third International Conference on Machine Learning.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. CoRR, abs/2506.11763.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. 2025. Mmreview: A multidisciplinary and multimodal benchmark for llm-based peer review automation. CoRR, abs/2508.14146.

Google. 2024. Gemini deep research — your personal research assistant. https://gemini.google/overview/deep-research/. Accessed: 2026-01-09.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. 2025. Is your LLM secretly a world model of the internet? model-based planning for web agents. Trans. Mach. Learn. Res., 2025.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 6864–6890. Association for Computational Linguistics.

Mengze Hong, Di Jiang, Weiwei Zhao, Yawen Li, Yihang Wang, Xinyuan Luo, Yanjie Sun, and Chen Jason Zhang. 2025. Multimodal peer review simulation with actionable to-do recommendations for community-aware manuscript revisions. CoRR, abs/2511.10902.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with LLM agents. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pages 1208–1226. Association for Computational Linguistics.

Xiaopeng Ke, Hexuan Deng, Xuebo Liu, Jun Rao, Zhenxi Song, Jun Yu, and Min Zhang. 2025. AQuilt: Weaving logic and self-inspection into low-cost, high-relevance data synthesis for specialist LLMs. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 15771–15798. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025a. Evaluating scoring bias in llm-as-a-judge. CoRR, abs/2506.22316.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. CoRR, abs/2504.21776.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. CoRR, abs/2212.10511.

Moonshot AI. 2024. Moonshot ai developer documentation: Use web search. https://platform.moonshot.ai/docs/guide/use-web-search. Accessed: 2026-01-08.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. CoRR, abs/2112.09332.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

OpenAI. 2025. Introducing openai deep research. https://openai.com/index/introducing-deep-research/. Accessed: 2026-01-05.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA, pages 311–318. ACL.

Perplexity. 2025. Introducing Perplexity Deep Research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research. Accessed: 2026-03-04.

Jun Rao, Xuebo Liu, Hexuan Deng, Zepeng Lin, Zixiong Yu, Jiansheng Wei, Xiaojun Meng, and Min Zhang. 2025. Dynamic Sampling that Adapts: Iterative DPO for Self-Aware Mathematical Reasoning. arXiv preprint arXiv:2505.16176.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS

2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Simra Shahid, Marissa Radensky, Raymond Fok, Pao Siangliulue, Daniel S. Weld, and Tom Hope. 2025. Literature-grounded novelty assessment of scientific ideas. CoRR, abs/2506.22026.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025. Deep research: A systematic survey. CoRR, abs/2512.02038.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as A multi-turn and long-context dialogue with role-based interactions. CoRR, abs/2406.05688.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. CoRR, abs/2312.11805.

Chunhao Tian, Yutong Wang, Xuebo Liu, Zhexuan Wang, Liang Ding, Miao Zhang, and Min Zhang. 2025. AgentInit: Initializing LLM-based multi-agent systems via diversity and expertise orchestration for effective and efficient collaboration. In Findings of the Association for Computational Linguistics: EMNLP 2025.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9440–9450. Association for Computational Linguistics.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020, pages 384–397. Association for Computational Linguistics.

Zhexuan Wang, Xuebo Liu, Li Wang, Zifei Shan, Yutong Wang, Zhenxi Song, and Min Zhang. 2026. MASPO: Joint prompt optimization for LLM-based multi-agent systems. In Proceedings of the 43rd International Conference on Machine Learning.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net.

xAI. 2025. Grok 3. https://x.ai/blog/grok-3. Accessed: 2026-03-04.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. 2025. Reviewrl: Towards automated scientific review with RL. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025, pages 16931–16943. Association for Computational Linguistics.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, and 4 others. 2026. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. CoRR, abs/2601.01576.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025a. Deep research: A survey of autonomous research agents. CoRR, abs/2508.12752.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. CoRR, abs/2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 29330–29355. Association for Computational Linguistics.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. Inf. Fusion, 124:103332.

A Appendix

A.1 Case Study

Although quantitative evaluation captures broad differences across systems, illustrating how our framework avoids generic statements and whether the generated reports approach the phrasing of human experts is often difficult to demonstrate directly through quantitative metrics alone. Therefore, we provide specific qualitative examples in Table 7 for supplementary explanation.

Reducing generic comparisons. The point-wise analysis process helps reduce generic, high-level paraphrasing and encourages the model to recover concrete technical details from the paper and related work. As shown in the first type of case study in Table 7, GPT-5 DeepResearch describe a related medical AI system merely as one that "simulated patient-doctor conversations," whereas our report identifies the more specific OSCE-style consultation setting and connects it to history-taking, diagnostic reasoning, and communication skill evaluation. Similarly, for mathematically oriented papers, our framework does not merely state that a paper uses "topology and category theory," but successfully identifies the first derived functor, quiver-representation resolutions, and the resulting path-count formulation. These examples fully demonstrate that decomposed novelty analysis can encompass more technical details, thereby effectively avoiding generic statements.

Memory-enabled expert-like judgment. Human reviews provided by historical memory serve as a reference, enabling the model's analysis to more closely align with a human expert's perspective, particularly in accurately defining the scope and level of the proposed contributions. As demonstrated in the second type of case study, while the baseline model without memory often provides generic praise or broad endorsements, our memory-augmented model can explicitly point out the exact nature of a contribution. For instance, in the clinical evaluation case, the report positions the work as an "integrative and domain-specialized" advancement rather than a "transformative methodological breakthrough." Furthermore, in the concept bottleneck model (CBM) case, the report notes that the individual building blocks are not unprecedented, but their systematic adaptation to the CBM framework is what makes it valuable. These calibrated expressions help precisely define the scope and level of the innovations, making the phrasing much closer to the professional standard of real human reviewers.

A.2 Historical Research Memory Implementation and Example

This appendix mainly presents a concrete example of the historical research memory, specifically illustrating the generated theme tree. Rather than repeating the construction pipeline already described in the main text, Figure 3 provides one produced memory branch for the theme *Search-Guided Planning and Optimization*. The figure should be read as a data-driven development tree over extracted novelty points: the top node is a broad theme, intermediate nodes are subthemes produced by clustering and LLM-based refinement, and terminal leaves are either fine-grained subthemes or individual papers attached to a sufficiently small theme.

In this example, node 3 splits into several research paths. Branch 3.1, for instance, is further divided into 3.1.1–3.1.4 because its clustered innovation points still contain multiple distinguishable directions, such as learned search for structured generation, neural-guided symbolic search, tree search for LLM planning and decoding, and factorized goal-search planning. By contrast, branch 3.4 directly lists papers because the branch itself contains fewer than 15 innovation records under the refinement rule, so the construction pipeline does not split it into additional child themes. This mixed structure is expected: the memory tree is not a manually balanced taxonomy, but a hierarchy whose depth reflects the local density and diversity of historically extracted innovation points.

To facilitate public access and exploration, the entire historical memory database is deployed as an interactive website, which is provided alongside our open-source code. Users can leverage this web interface to intuitively navigate the global memory structure. For example, users can utilize keyword searches to locate a specific node within the database and examine its adjacent nodes. Additionally, they can view the specific development history of the current node itself, or explore the broader evolutionary trajectory of its parent aggregated nodes.

A.3 Cross-Venue Breakdown

Table 8 provides the detailed scores broken down by each conference or journal for the cross-venue generalization experiment. Across these diverse

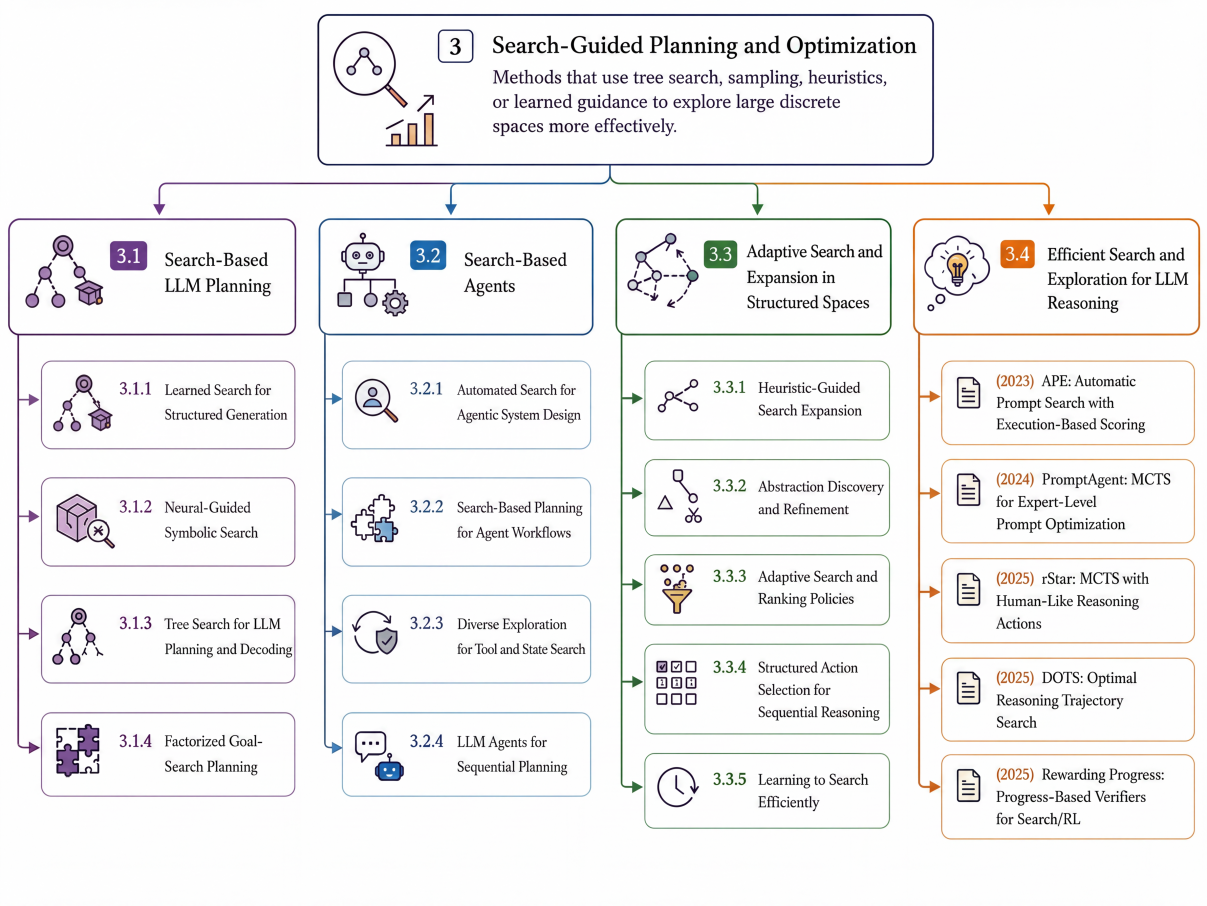


Figure 3: An example of theme tree from the historical research memory: a clustered development tree for the search-guided planning and optimization theme. Branches are recursively refined only when the corresponding theme exceeds the refinement threshold; therefore, some branches are further divided into subthemes, while others directly list papers.

venues (ACL, ACM MM, CVPR, JMLR, and TPAMI), MemoNoveltyAgent demonstrates solid overall results and remains the best system in every venue. The margin is not identical across venues, but the pattern is stable: MemoNoveltyAgent improves completeness, depth, and faithfulness in most venue-level slices, while fluency remains comparable to strong baselines.

Because the historical memory is constructed based on ICLR papers, directly transferring it to these other venues causes a slight decrease in the absolute scores. Nevertheless, MemoNoveltyAgent still consistently outperforms other baselines, maintaining a clear advantage.

A.4 Database and Retrieval Details

To construct the local literature database, we expand the target paper's first-order references to include second-order citations. To manage the noise and scale of this expanded set, we prioritize papers based on structural relevance (i.e., co-occurrence among first-order references) and recency. This targeted selection is based on a reasonable assumption: recent papers with high local exposure in a specific research neighborhood are more likely to capture its core directions and to overlap with the target paper in novelty. They therefore constitute the most important potential sources of novelty conflict. Furthermore, we incorporate the full-text content of these prioritized papers rather than relying solely on abstracts. This allows the agent to access intricate technical details, thereby facilitating a significantly deeper and more granular analysis of potential novelty conflicts.

Following PDF parsing, we explicitly remove reference lists to minimize noise. The cleaned text is then divided into chunks of 512 tokens. Each chunk is encoded into a dense vector representation using the *maidalun1020/bce-embedding-base_v1* model and indexed in a vector database. During the retrieval phase, the LLM generates targeted search queries for each novelty point, dynamically

tailored to specific task requirements and contextual nuances. To ensure comprehensive coverage, exactly 6 queries are generated per novelty point: 4 queries composed of relevant keywords and 2 queries formulated as specific contextual sentences. Inspired by RAGFlow[1], we adopt a two-stage hybrid retrieval approach. The first stage conducts initial recall through two parallel pathways: a sparse retrieval path based on BM25 lexical matching to capture exact terms (e.g., method names), and a dense retrieval path based on vector similarity to capture semantic relevance. The scores from both pathways are normalized and fused. In the second stage, the top-ranked candidates from the hybrid recall are forwarded to the Qwen3-Reranker-4B cross-encoder model (Zhang et al., 2025b) for fine-grained relevance estimation. Finally, the reranker selects the top 7 most relevant chunks for each query, which are subsequently provided to the Analyst Agent—and the checklist evaluator—as highly relevant, citation-grounded evidence.

A.5 Citation Accuracy

For Citation Accuracy (CA), we follow the core idea of DeepResearchBench (Du et al., 2025): citation-bearing claims are extracted from the report, mapped to the cited full-text papers, and judged according to whether the cited source actually supports the claim. Under this protocol, self-validation improves CA from 89.90 to 93.72, while GPT-5 DeepResearch obtains 87.60. This indicates that the Validator and Improver agents do not merely polish the report, but materially reduce unsupported or misattributed citation claims.

A.6 System Efficiency and Cost Analysis

Whenever a new target paper is evaluated, the system needs to crawl data to construct a tailored reference dataset. The exact speed of this process depends heavily on the network conditions of the environment running the agent system. To evaluate the system's efficiency, we selected 10 papers to measure the time cost. The primary time consumption in our pipeline comes from the chunking and embedding of the reference literature, as well as the RAG retrieval process. The remaining time consists of LLM API calls, which can be easily and significantly accelerated through parallelization.

On average, for each paper, the RAG retrieval process takes approximately 30.5 seconds, while the chunking and embedding process takes about 701.3 seconds. For comparison, the GPT series' deep research baseline in our study takes an average of about 1100 seconds. Furthermore, it occasionally suffers from stalling issues, requiring up to 1 hour and 15 minutes to resolve in such edge cases. In contrast, baselines relying solely on simple web search are much faster, taking only about 25 seconds.

In terms of API cost, our work is highly economical. Because we utilize a smaller, cost-effective model (GPT-5-mini) and perform comparisons based on RAG-retrieved content (i.e., comparing relevant retrieved text chunks rather than feeding the entire full texts into the context window), the average cost is only $0.075 per paper according to OpenAI's API pricing. In contrast, the GPT series' deep research baseline is expected to incur a significantly higher cost, although it is difficult to calculate its exact token consumption directly.

Overall, our approach strategically saves on expensive resources, such as API calls and GPU compute, by relying primarily on much cheaper resources like network bandwidth and CPU processing. While this trade-off may result in a longer processing time per paper, the overall financial cost remains relatively low. Furthermore, since these network and CPU-bound processes are highly parallelizable, our system can be easily scaled up for efficient, large-scale execution.

Additionally, in our open-source version, users can easily adjust the size of the reference database for each paper (the default is set to 200 related papers). If a user reduces this number to 50, the database construction time drops to approximately 180 seconds, which can effectively improve the overall efficiency of the system based on user needs.

[1] https://github.com/infiniflow/ragflow

Case Type 1: Generic Statements vs. Fine-Grained Technical Details	
Context	Evaluating the specific technical alignment and similarities with AMIE.
GPT-5 DeepResearch	AMIE also simulated patient-doctor conversations to improve diagnosis. *Overly generic*
MemoNoveltyAgent	Resembling the proposed approach, AMIE utilized **OSCE-style simulated consultations** for training and evaluation to **assess history-taking, diagnostic reasoning, and communication skills.**
Context	Evaluating whether the report explains the mathematical mechanism behind a proposed emergence metric.
GPT-5 DeepResearch	The paper provides a formal measure of emergence using topology and category theory. *High-level restatement*
MemoNoveltyAgent	The paper identifies emergence with the **first derived functor** of an observation functor, computes it through **projective/injective resolutions in quiver representations**, and reduces the result to a **path-count formula** over affected network edges.
Case Type 2: Monolithic Paraphrasing vs. Calibrated Expert Judgment	
Context	Evaluating the nature and scope of a paper's core contributions.
Baseline *(w/o Memory)*	The manuscript presents a well-motivated, non-obvious integration that meaningfully advances multimodal, sequential clinical evaluation. *Generic praise*
MemoNoveltyAgent	The component techniques are adapted from existing agent work, so the contribution is **integrative and domain-specialized** rather than a **transformative methodological breakthrough.**
Context	Evaluating a deployment-oriented LLM routing and abstention framework.
Baseline *(w/o Memory)*	The system-level design and Pareto analysis constitute a substantial and distinctive contribution. *Mostly affirmative*
MemoNoveltyAgent	The framework is **deployment-ready and operationally valuable**, but its individual components mainly **refine rather than revolutionize** selective evaluation, calibration, and model-routing literature.
Context	Evaluating the originality of an editable concept bottleneck model framework.
Baseline *(w/o Memory)*	The three-level taxonomy and influence-function formulation are treated as genuine methodological contributions. *Less calibrated*
MemoNoveltyAgent	The individual building blocks are not unprecedented, but the **systematic adaptation to concept bottleneck models is valuable** because it organizes editing, correction, and concept-level intervention into a coherent CBM-specific framework.
Context	Evaluating whether a new sequence model architecture is a breakthrough or a targeted synthesis.
Baseline *(w/o Memory)*	The architecture is described as a substantial design-level contribution. *Broad endorsement*
MemoNoveltyAgent	The contribution is better understood as a **targeted architectural synthesis** rather than a wholesale departure, since it builds on prior Mamba, Liquid, S3, and SPACETIME-style ideas while adapting them to the paper's specific modeling objective.
Context	Evaluating the contribution of a modified Transformer attention operator.
Baseline *(w/o Memory)*	The report presents the method as a cohesive algorithmic, theoretical, and empirical contribution. *Less discriminating*
MemoNoveltyAgent	The core novelty is a **concrete operator-level change** in the attention mechanism, while the accompanying stability techniques are mostly **supporting engineering choices** rather than independent conceptual breakthroughs.

Table 7: Case studies illustrating report quality. Case Type 1 shows that point-wise analysis helps recover concrete technical details rather than generic comparisons. Case Type 2 shows that historical research memory helps calibrate novelty judgments by distinguishing genuine contributions from domain-specific integrations, targeted refinements, or supporting engineering choices.

Venue	Dimension	DeepReview	MemoNoveltyAgent	GPT-5.4 DeepResearch	GPT-5.4 Online
ACL	Completeness	8.61	9.07	7.84	6.00
	Depth	7.07	8.15	7.74	3.48
	Effectiveness	9.77	9.94	9.82	9.59
	Faithfulness	6.48	7.93	6.79	5.79
	Fluency	9.86	9.00	9.50	9.79
	Overall	8.36	8.82	8.34	6.93
ACM MM	Completeness	8.65	9.42	8.10	6.23
	Depth	7.96	9.04	7.89	3.41
	Effectiveness	9.82	10.00	9.47	9.47
	Faithfulness	7.59	8.48	6.83	5.31
	Fluency	9.21	9.86	9.71	9.71
	Overall	8.65	9.36	8.40	6.83
CVPR	Completeness	8.58	8.74	7.52	6.55
	Depth	8.04	8.70	7.59	3.96
	Effectiveness	9.82	10.00	9.65	9.71
	Faithfulness	7.24	9.10	6.72	5.69
	Fluency	9.50	9.57	9.43	9.71
	Overall	8.64	9.22	8.18	7.12
JMLR	Completeness	7.84	8.16	7.42	6.58
	Depth	7.44	7.93	7.07	4.63
	Effectiveness	9.71	10.00	9.77	9.47
	Faithfulness	7.41	9.00	7.35	3.86
	Fluency	9.29	9.71	9.00	9.86
	Overall	8.34	8.96	8.12	6.88
TPAMI	Completeness	8.42	8.71	7.32	6.16
	Depth	8.07	8.67	6.78	3.85
	Effectiveness	9.82	9.77	9.59	9.24
	Faithfulness	7.79	8.31	6.86	5.24
	Fluency	9.86	9.29	9.36	9.57
	Overall	8.79	8.95	7.98	6.81

Table 8: Detailed venue-level breakdown for cross-venue evaluation (0–10 scale).

Dimension	Better (Ours)	Worse (Ours)	Tie
Completeness	100.0%	0.0%	0.0%
Depth	100.0%	0.0%	0.0%
Effectiveness	100.0%	0.0%	0.0%
Faithfulness	100.0%	0.0%	0.0%
Fluency	100.0%	0.0%	0.0%
Overall Win Rate	**100.0%**	**0.0%**	**0.0%**

Table 9: Pairwise comparison results between our generated reports and OpenNovelty reports on 10 randomly selected ICLR 2026 submissions. “Better” indicates the percentage of cases where our report was preferred.

A.7 Comparison with OpenNovelty

OpenNovelty(Zhang et al., 2026) is highly relevant to our research, as it also conducts in-depth investigations into the novelty of academic papers. However, it cannot be functionally reproduced at this time because its open-source repository relies on a private API that has not yet been publicly released. To conduct a comparison, we randomly selected 10 reports generated for ICLR 2026 submissions from their publicly available examples and compared their quality with our generated reports.

To minimize the bias caused by the structural differences between our reports and theirs, we adopted a pairwise comparison method following previous work(Zheng et al., 2023). Furthermore, we utilized the same scoring criteria provided to our human judges as the rubric to guide the LLM's assessment. As shown in Table 9, our reports outperformed OpenNovelty across all evaluated dimensions.

However, for the sake of fairness, we must point out that the core objective of OpenNovelty is fundamentally different from ours. OpenNovelty focuses on making a binary (yes or no) judgment on whether a proposed innovation conflicts with prior work, providing clear, traceable evidence to assist human reviewers, without involving further AI-driven analysis or reasoning. Although its results are presented in the form of a report, the comprehensive quality of the report text is not its absolute core focus. In contrast, our evaluation criteria are primarily based on the overall quality of the generated reports, which explains why our results are completely superior to theirs under this specific evaluation framework.

A.8 Output Report Example

The following presents an output example of the AgentClinic report. Due to space limitations, this appendix only provides the complete Section 1, a representative novelty point from Section 2, and selected excerpts from Section 3. Omitted portions are explicitly marked.

A.9 Human Evaluation Instructions

To validate the automated evaluation method, we asked experienced Master's students specializing in artificial intelligence to score generated reports on a 1–5 scale. The rubric below translates the manual scoring guide used by annotators and keeps it in the same full-width framed format as the prompt examples.

A.10 Evaluation Checklist and Core Prompts

The following figures present the core prompts used in the evaluation and report-generation pipelines. For readability, repetitive instructions are omitted, while the task definitions, input fields, constraints, and output formats are retained.

Evaluation Checklist Items. The implemented checklist contains 118 yes/no questions: 14 for Fluency, 17 for Effectiveness, 31 for Completeness, 29 for Faithfulness, and 27 for Depth.

Output Report Example: Paper Content Summary

1. Paper Content Summary

This paper introduces AgentClinic, a publicly released, multimodal agent benchmark designed to evaluate large language (and vision) models in simulated clinical environments that require interactive, dialogue-driven, sequential decision making rather than static question-answering. It aims to address the gap that conventional medical benchmarks (e.g., MedQA/USMLE) present all information up-front and do not capture iterative history-taking, test-ordering, multimodal image interpretation, tool use, bias effects, multilingual needs, specialist domains, or patient-centered outcomes. To do so, the authors implement four LLM-based agents (doctor, patient, measurement/instrument reader, and moderator) and populate cases from USMLE/MedQA, deidentified EHRs (MIMIC-IV), NEJM case challenges, and specialist datasets (MedMCQA). The benchmark supports nine specialties, seven languages, and 23 cognitive/implicit bias scenarios injected via agent role prompts. Doctor agents can use an "agent toolbox" of capabilities including chain-of-thought prompting (zero/one-shot/reflection), adaptive retrieval-augmented generation (RAG) from textbooks or web sources, an experience-persistent notebook for experiential learning, and multimodal image requests.

Evaluation compares multiple LLM backbones (Claude-3.5, GPT-4/4o/3.5, Mixtral-8x7B, Llama variants) across AgentClinic-MedQA, AgentClinic-MIMIC-IV, and multimodal NEJM cases using metrics beyond diagnostic accuracy. These include patient-agent confidence, compliance, consultation willingness, and clinician dialogue realism/empathy ratings. Key findings reported are that sequential, interactive diagnosis is substantially harder than static MedQA (diagnostic accuracy can fall dramatically and MedQA performance is only weakly predictive of AgentClinic performance), Claude-3.5 often attains the highest accuracy across settings, models differ markedly in their ability to benefit from tools (the notebook yields large gains for some models such as Llama-3), GPT-4 shows relative robustness to instructed biases while models like Mixtral degrade more, and simulated biases can reduce patient-agent trust and compliance.

Definitions: Objective Structured Clinical Examination (OSCE) is a dialogue-driven clinical skills exam template used to structure cases; Retrieval-Augmented Generation (RAG) denotes retrieving external documents to ground model generations; Sparse Mixture-of-Experts (SMoE) is an architecture that routes tokens to a subset of expert blocks (used by Mixtral); USMLE is the United States Medical Licensing Exam; NEJM refers to the New England Journal of Medicine.

Figure 4: Paper content summary excerpted from the AgentClinic report.

Output Report Example: Point-wise Novelty Analysis

2.1. Novelty Point 1: (Classification: Dataset/Benchmark)
a) Claimed Novelty. Explanation: AgentClinic is presented as an open-source dataset/benchmark suite that simulates longitudinal, dialogue-driven clinical encounters in an OSCE-style framework, extending static medical QA into interactive, multimodal clinical workflows. Concretely, AgentClinic composes case JSONs (patient/doctor/measurement/moderator agents and persistent state) by drawing on USMLE/MedQA items, NEJM case challenges, and de-identified EHR records (MIMIC-IV). The benchmark explicitly supports multimodal inputs (clinical text plus medical images), interactive actions (ordering/requesting tests, invoking image reads, performing measurements), persistent documentation (notes that carry across the encounter), and patient-centric follow-ups (confidence, compliance, consultation ratings). It is organized into variants (AgentClinic-MedQA, -MIMIC-IV, -NEJM, -Spec, -Lang) covering nine specialties and seven languages, and the paper reports large-scale experiments across seven LLM backbones, example model accuracies (e.g., Claude-3.5 ~62.1% +/-3.3 on AgentClinic-MedQA), multimodal NEJM evaluations, specialty/language breakdowns, and clinician-rated realism/empathy scores – demonstrating both the resource and its empirical utility.
b) Similarities. - The cited AMIE work presents an approach that develops conversational diagnostic AI using simulated OSCE-style interactions: it builds a simulated clinician/patient role-play and self-play simulated-dialogue pipeline to train and evaluate diagnostic dialogue systems, and it evaluates AMIE via randomized OSCE-style text consultations with specialist physician and patient-actor assessments. [1] - Use of simulated patient-physician conversations and OSCE-like data-generation procedures resembles earlier datasets and work that create simulated clinical dialogues for training/evaluation. [2] - Building on existing medical QA benchmarks (USMLE/MedQA/MedMCQA) as source material follows established practice in medical-LM evaluation. [3] ...
c) Unique Differences. - AgentClinic combines multimodal clinical inputs (text + medical images) with an OSCE-style, sequential, tool-enabled clinical workflow in a single open benchmark – prior medical OSCE-style evaluations were predominantly text/chat-based and prior multimodal medical LLM work rarely packaged into an interactive clinical-encounter benchmark. [1] [4] - The benchmark's explicit case JSON schema that separates patient/doctor/measurement/moderator agents, supports persistent notes and measurements, and exposes explicit test-ordering and image-read actions (i.e., structured tool-like clinical operations) is not matched by earlier medical dialogue datasets or by general agent benchmarks that are not specialized for clinical semantics. [2] [6] ...
d) Details of Unique Differences. As a dataset/benchmark, AgentClinic fills a concrete gap: prior OSCE-style medical evaluation work demonstrated that LLM-based systems can be assessed on multi-turn diagnostic dialogue with clinician and patient ratings, but those studies were largely text-only and focused on diagnostic accuracy and communication metrics within simulated dialogue systems; they did not simultaneously provide a standardized, multimodal, tool-enabled benchmark that (1) encodes clinical actions (test orders, image reads, persistent EHR-like notes) as first-class operations, (2) sources cases from both standardized exam items (USMLE/MedQA) and high-complexity clinical narratives (NEJM) plus real EHR data (MIMIC-IV), and (3) includes multilingual and multi-specialty partitions with patient-centered follow-ups. The practical consequence is that AgentClinic enables evaluation of capabilities that matter for deployed clinical agents – sequential decision-making under information-gathering constraints, multimodal image+text interpretation in the loop, persistent documentation and state, and patient-facing metrics – all in one reproducible suite. This combination meaningfully extends prior diagnostic-dialogue and multimodal-medical-LM work by converting isolated capabilities into an integrated, clinically-grounded benchmark that can measure whether models coordinate information acquisition, tool use, and communication across an encounter – a capability set that earlier medical benchmarks and general agent suites either tested separately or did not operationalize for clinical semantics. [1] [4] [5]
Additional novelty points in the source report are omitted only for space.

Figure 5: Point-wise novelty analysis excerpted from the AgentClinic report.

Output Report Example: Novelty Summary

3. Novelty Summary
Overall characterization. - This work assembles an integrated, reproducible clinical agent-evaluation stack that moves beyond isolated static medical QA by combining OSCE-style multi-turn encounters, multimodal inputs (images + text), explicit tool-like clinical operations (test ordering, image reads, persistent documentation), agent-role specialization, and configurable bias injections – and uses that stack to run controlled, factorial experiments measuring diagnostic accuracy, tool effects, bias provenance, and patient-centred outcomes. The core contribution is an applied, domain-specialized evaluation and experimental platform that operationalizes several agent-evaluation desiderata previously explored separately into a single clinical workflow.
Genuine innovations (excerpted). - What is new and how it is realized: The benchmark packages clinical-case JSON templates that explicitly encode patient/doctor/measurement/moderator agents, persistent encounter state (notes, measurements), requestable diagnostic instruments, and patient follow-ups, while sourcing cases across exam-style items, high-complexity narrative cases, and de-identified EHR rows and offering multilingual and specialty partitions. It therefore converts static medical QA items into reproducible, sequential clinical tasks with multimodal stimuli and downstream patient-impact signals. ... - What is new and how it is realized: The system treats measurement instruments as first-class agents (request -> instrument output -> clinician interpretation), enforces role-specific information exposure and constrained inquiry budgets, provides runtime tool access (adaptive retrieval, CoT/reflection variants), and implements a persistent notebook that carries experience across cases – all evaluated and ablated empirically. The measurement-as-agent and persistent cross-case notebook combined with clinical inquiry budgets instantiate a realistic diagnostic test loop that prior general multi-agent frameworks did not operationalize for clinical semantics at scale. ...
Aspects lacking substantial originality (excerpted). - Use of OSCE-style simulated dialogues and self-play simulated patient/clinician role-play – Category: Standard Practice. Improvement suggestion: incorporate prospectively collected clinician–patient recorded interactions or clinician-in-the-loop adjudication to further validate and diversify dialogue dynamics. ... - Sourcing questions/cases from exam items (USMLE/MedQA) and NEJM case challenges – Category: Standard Practice. Improvement suggestion: add protocols and leakage analyses to guarantee no contamination with model-pretraining corpora and prioritize clinician-authored novel vignettes for higher ecological novelty. ...
Final One-line Summary. 3 - Good / Significant: This paper delivers a significant, well-executed advance by packaging multimodal, tool-enabled OSCE-style clinical evaluation, a domain-specialized multi-agent simulation (including measurement-as-agent and persistent notebook), and a clinically focused bias-injection and empirical study suite – providing an important resource and actionable empirical findings for clinical AI assessment – while most component techniques are adapted from existing agent and bias-evaluation work so the contribution is integrative and domain-specialized rather than a transformative methodological breakthrough.

Figure 6: Novelty summary excerpted from the AgentClinic report.

Human Evaluation Rubric: Fluency, Effectiveness, and Completeness

Fluency. Core criteria: grammatical correctness, concise expression, readability, and absence of repetition. **5 - Excellent:** The report is extremely fluent; all sentences are grammatically correct; capitalization is proper; the content is concise; there is no cross-section repetition or redundant restatement of technical details; terminology is consistent; and the reading experience is excellent. **4 - Good:** The report is generally fluent and grammatical, with at most isolated redundant expressions or slightly verbose sentences; the overall structure is compact and terminology is mostly consistent. **3 - Fair:** The report is readable but contains visible verbosity or repetition, and may contain minor grammar/spelling errors or overly complex sentences. **2 - Poor:** The report is difficult to read because of multiple grammar or formatting issues, repeated content, improper wording, or confused terminology. **1 - Very Poor:** The report is almost unreadable, logically confused, filler-heavy, and not aligned with academic-report writing standards.

Effectiveness. Core criteria: strict three-section template compliance, logical coherence, and no irrelevant content. **5 - Excellent:** The report strictly follows the three-section structure of content summary, comparative analysis, and novelty summary, without extra headings or sections; Section 2 contains the required a/b/c/d items with coherent logic; Section 3 is fully grounded in earlier analysis and includes the required 1–4 score and one-sentence summary; the report stays strictly on the assigned paper. **4 - Good:** The report follows the three-section structure and satisfies the main formatting requirements, with at most minor subsection-title deviations or a slightly abrupt Section 3 summary. **3 - Fair:** The overall structure is roughly correct but has visible format violations, such as a missing required subsection or a broken logic chain. **2 - Poor:** The structure is confused, the three main parts are not clearly separated, or the report contains substantial irrelevant domain history/background; the required score or summary is missing. **1 - Very Poor:** The report ignores the template or becomes a generic essay unrelated to the assigned paper-analysis task.

Completeness. Core criteria: coverage of the main paper's core elements, all novelty points, and related-paper comparisons. **5 - Excellent:** There are no meaningful omissions; the summary includes the target paper's objective, methods, and uncommon abbreviations; the comparative analysis identifies all similar innovation points mentioned in related literature and fully analyzes similarities and unique differences for every main point; the reported differences are genuinely unique. **4 - Good:** The report covers most core content and main novelty points, but some secondary comparison details may be insufficient or some similarities may be missed. **3 - Fair:** The report has clear omissions, such as missing a key method, omitting an important related paper, or claiming a difference that has an obvious similar prior work. **2 - Poor:** The report misses key parts, ignores major novelty points, omits many similar papers, or gives non-unique "unique" differences. **1 - Very Poor:** The report provides very little information, omits most retrieved evidence, and gives insufficient analysis.

Figure 7: Human evaluation rubric, part 1.

Human Evaluation Rubric: Faithfulness and Depth

Faithfulness. Core criteria: grounding in retrieved text, no hallucination, correct attribution, and no exaggeration. **5 - Excellent:** The content is entirely based on the provided target-paper and related-paper text; all similarities and differences have clear textual evidence; the report does not exaggerate achievements; all factual claims are grounded; and there is no misunderstanding or contradiction. **4 - Good:** Most content is accurate, with at most slight ambiguity in a secondary detail and no core factual error or serious hallucination. **3 - Fair:** The report contains questionable statements, partial misunderstanding, distorted meaning, or claims that are only inferred rather than directly supported. **2 - Poor:** The report contains clear hallucinations or errors, such as fabricated data/comparisons or attributing related-paper content to the target paper. **1 - Very Poor:** The report contradicts or is unrelated to the retrieved text and contains unsupported speculation or fabrication.

Depth. Core criteria: concrete technical terminology, quantitative or specific evidence, and independent critical thinking. **5 - Excellent:** The analysis is highly detailed, uses specific names rather than vague descriptions, cites concrete results to support claims, and shows independent thinking by identifying limitations or uncertainty and giving evidence-based novelty judgments. **4 - Good:** The report includes major technical terms and some concrete comparisons and can identify mechanism-level differences, but may lack quantitative support; the analysis is mostly independent. **3 - Fair:** The report remains at the level of functional description, says that something is better or different without explaining why in technical detail, and mostly restates the original content. **2 - Poor:** The report is superficial, uses vague promotional adjectives without technical support, largely copies author claims, and lacks independent judgment. **1 - Very Poor:** The report is empty or vague, misses the core technical contribution, cannot reason independently, and may contradict earlier analysis.

Figure 8: Human evaluation rubric, part 2.

Evaluation Checklist: Fluency (14 questions)

1. Is the report free from formatting issues and correctly capitalized throughout? **2.** Are all sentences grammatically correct and free from errors? **3.** Are verb tenses used consistently within each section? **4.** Do pronouns agree in number and gender with their antecedents? **5.** Does the report avoid unnecessary repetition across sections when that repetition does not add comparison, evidence, or synthesis? **6.** Does the report avoid elaborations that are tangential to the paper or its novelty evaluation? **7.** When technical details recur across sections, are they used for a new analytical purpose rather than copied verbatim? **8.** Is the language precise enough for academic evaluation while preserving necessary technical and comparative detail? **9.** Is the report easy to read, without unnecessary complexity? **10.** Are sentences generally concise, avoiding unnecessary wordiness? **11.** Is terminology used consistently to avoid reader confusion? **12.** Are paragraphs organized in a way that makes the argument easy to follow? **13.** Does the report maintain a clear academic tone throughout? **14.** Are section and subsection labels clear enough for readers to locate the relevant analysis quickly?

Figure 9: Fluency checklist questions used in checklist evaluation.

Evaluation Checklist: Effectiveness (17 questions)

1. Is the current report strictly divided into sections as required: 1. Paper Content Summary, 2. Point-wise Novelty Analysis, 3. Novelty Summary **2.** Are there no extra sections or headers beyond the three required sections? **3.** Does the 'Novelty Summary' (Section 3) draw its conclusions logically from the evidence and analysis presented in the preceding sections? **4.** Does Section 3 clearly connect its final judgment to the point-wise analyses in Section 2? **5.** In the 'Point-wise Novelty Analysis' (Section 2), does the analysis in subsection 2.d ('Details of Unique Differences') logically follow from and focus exclusively on the unique aspects identified in its corresponding subsection 2.c? **6.** Does the analysis within subsection 2.d ('Details of Unique Differences') directly address the required questions for the specific innovation 'Classification' defined in subsection 2.a? **7.** Does Section 2 maintain a clear organization between identified novelty points and their corresponding comparative analyses? **8.** Within Section 2, does the report correctly use the required subsection headers: 'a. Claimed Novelty', 'b. Similarities', 'c. Unique Differences', and 'd. Details of Unique Differences' (if applicable)? **9.** Does each sub-section under each section contain the required content in the template? **10.** Does Section 3 end with a single-line summary of the innovation impact? **11.** In Section 3, was the article's originality rated on a scale of 1 to 4 as required by the template? **12.** Does the report clearly distinguish genuine contributions from less original or incremental aspects when making the final novelty judgment? **13.** Does the final one-line summary include both an innovation score/rating and a concise explanation of the main strength and limitation? **14.** Does each piece of content strictly adhere to the proposed theme without any signs of deviation? **15.** When broader domain history or related-work background is included, is it directly relevant to evaluating the paper's novelty rather than tangential? **16.** Does the entire report remain focused on analyzing the specified paper without introducing irrelevant topics or tangential discussions? **17.** Does every section of the report address its designated purpose (content summary, point-wise novelty analysis, or novelty evaluation) without mixing or conflating objectives?

Figure 10: Effectiveness checklist questions used in checklist evaluation.

Evaluation Checklist: Completeness (31 questions)

1. Does the Paper Content Summary include the main objectives, methodology, and abbreviations that are mentioned in the provided text from the Main paper? **2.** Does the summary explicitly state the paper's main objective in a clear, single sentence, if this objective is mentioned in the provided text? **3.** Does the summary restate the specific problems the Main paper claims to solve, as described in the provided text? **4.** Does the summary enumerate the key methodological components that are described in the provided text from the Main paper? **5.** Does the summary define the uncommon abbreviations that appear in the provided text from the Main paper? **6.** Does the summary include the main paper's most important technical components without drifting into unrelated background? **7.** In the Point-wise Novelty Analysis, does the report identify the main claimed novelty points explicitly mentioned in the retrieved text of the main paper? **8.** Does the Point-wise Novelty Analysis section avoid overlooking central novelty points from the retrieved text or combining distinct major contributions in a way that obscures the evaluation? **9.** Does the report include the main theoretical or conceptual contributions claimed by the Main paper when they are available? **10.** For theoretical novelty points, does the report cover the central theorem, proof target, formal claim, or conceptual framework described by the Main paper? **11.** In the similarity analysis section, does the report correctly identify similarities to the innovation point that are mentioned within the provided text from related papers? **12.** If the provided text from related papers describes prior methods with the same core algorithmic idea, does the report identify them as similar? **13.** If the provided text from related papers describes prior works sharing the same problem formulation or task, does the report identify them as similar? **14.** If the provided text from related papers describes alternative approaches to the same limitation or gap, does the report identify them as similar? **15.** Does the report cover similarities in datasets, benchmarks, tasks, or evaluation settings when the related-paper evidence supports them? **16.** Does the report cover similarities in empirical observations, trends, or diagnostic findings when the related-paper evidence supports them? **17.** Does the report cover related works that challenge the paper's claimed novelty, not only related works that support it? **18.** Does the report cover related works that are similar in theoretical framing or assumptions when such evidence is available? **19.** Does the report cover related works that are similar in empirical finding or diagnostic conclusion when such evidence is available? **20.** In the Unique Differences section, are the claims of uniqueness free from contradictions when checked against the provided text from related papers? **21.** Does the report avoid claiming a feature as a Unique Difference if the provided related-paper evidence describes a similar feature, even if different terminology is used? **22.** For each major unique point claimed in the report, does the report cover whether similar items are found within the provided text from related papers? **23.** After comparing with the provided text from related papers, does the report cover whether the uniqueness claims made in the report still appear to hold? **24.** Does the report cover whether the assertion of uniqueness for a specific feature is supported by the absence of that feature in the provided texts from related papers? **25.** Does the report cover differences in dataset scale, annotation, coverage, benchmark design, or evaluation method when those differences are central to novelty? **26.** Does the report cover cases where a dataset, benchmark, or system is valuable but follows an established design pattern? **27.** Does the report make clear which related papers support major unique-difference claims? **28.** Does the report cover standard engineering or evaluation practices when the paper presents them as novelty? **29.** Does the report avoid omitting a key limitation of prior work that the Main paper explicitly addresses? **30.** Does the report cover differences in theoretical assumptions, proof targets, or analytical scope when those differences are central to novelty? **31.** When a claimed difference is mainly a minor implementation change, simple combination, or standard extension, does the report explicitly reflect that limited originality?

Figure 11: Completeness checklist questions used in checklist evaluation.

Evaluation Checklist: Faithfulness (29 questions)

1. Does the 'Paper Content Summary' accurately reflect the paper's main task, objective, and methods, with all factual details being directly verifiable from the provided text of the Main paper? **2.** Does the 'Paper Content Summary' contain no information that is not explicitly supported by the provided text from the main paper? **3.** For each 'Claimed Innovation' subsection, do the stated classification, core idea, and key technical elements accurately represent the contributions as described in the provided text from the Main paper? **4.** Are all factual details within the 'Claimed Innovation' subsections directly verifiable from the provided text of the main paper? **5.** Does the report avoid describing a result as universal or general if the provided Main-paper evidence only supports a limited setting? **6.** Are all 'Similarities' (b) and 'Unique Differences' (c) described in the 'Comparative Analysis' section exclusively based on and directly supported by the content found within the provided text from related papers? **7.** Are all specific examples of similar objectives or mechanisms mentioned in the 'Similarities' subsections (b) precisely aligned with the information present in the provided text from related papers? **8.** Are all specific contrasts elaborated in the 'Unique Differences' subsections (c) directly and accurately supported by the provided text from related papers? **9.** Are the explanations and details provided for each 'Similarity' (b) and 'Unique Difference' (c) fully and accurately supported by the provided text from related papers? **10.** When the report says two works share a dataset, task, benchmark, or evaluation setting, is that shared setting supported by the provided related-paper evidence? **11.** When the report states that a related work reports a quantitative result, is the number or trend consistent with the provided evidence? **12.** Does the report avoid treating retrieved snippets as saying more than they actually say when drawing similarity or difference conclusions? **13.** Does the report avoid claiming a related work includes an ablation, experiment, proof, or benchmark that is not present in the provided evidence? **14.** Does the report avoid using references only as decoration when the surrounding sentence makes a substantive factual claim? **15.** When the report summarizes a cluster of related papers, does it avoid making the cluster sound more homogeneous than the evidence supports? **16.** When the report discusses prior limitations, does it tie those limitations to the correct related work or group of related works? **17.** Does the report avoid conflating multiple related papers into one source when attributing a method, result, or limitation? **18.** Does the report connect cited related works to the specific innovation point being discussed rather than using them as generic background? **19.** When the report discusses prior advantages, does it tie those advantages to the correct related work or group of related works? **20.** Does the report correctly distinguish between the Main paper's own contributions and the existing baselines or prior work described in the text, avoiding misattribution of features? **21.** Does the report accurately interpret the logical flow and causal relationships of the proposed method of the Main paper, ensuring that the mechanism is explained exactly as the authors intended without logic errors? **22.** Are the conclusions and claims in the report faithful to the nuances of the Main paper, avoiding exaggeration of results or misinterpretation of the limitations stated in the text? **23.** Are all 'Similarities' and 'Unique Differences' presented in the 'Comparative Analysis' section strictly derived from and explicitly supported by the content found within the provided text of related papers? **24.** Is the 'Details of Unique Differences' for each innovation point solely based on the unique aspects and details explicitly present in the provided text, without additional interpretation? **25.** Are any performance advantages or comparative benefits claimed within the report directly and explicitly supported by comparative evidence present in the provided text of related papers? **26.** Does the report avoid saying that prior work lacks a component when the provided evidence indicates that it has a comparable component? **27.** Does the report distinguish between a related paper's claimed contribution and the current report's evaluation of that contribution? **28.** When the report later makes an overall judgment about a contribution, is that judgment consistent with the evidence and conclusions already presented in the Unique Differences subsection? **29.** Are the interpretations in 'Details of Unique Differences' grounded in the identified unique aspects and connected to the available evidence, rather than introducing unrelated or unsupported claims?

Figure 12: Faithfulness checklist questions used in checklist evaluation.

Evaluation Checklist: Depth (27 questions)

1. Does the 'Paper Content Summary' include specific technical terminologies (e.g., specific model architectures, loss functions, mathematical formulations) rather than just high-level functional descriptions? **2.** Does the report explicitly define the specific input data types, datasets, or benchmark environments used in the main paper? **3.** When discussing empirical or application-oriented contributions, does the report specify the datasets, benchmarks, domains, or experimental settings involved rather than describing evaluation only generically? **4.** For dataset or benchmark contributions, does the report describe the resource's scale, composition, annotation, task coverage, or evaluation purpose when available? **5.** For empirical or analytical contributions, does the report identify the non-obvious finding rather than merely stating that experiments were conducted? **6.** Does the report connect claimed theoretical contributions to formal objects such as assumptions, propositions, theorems, proofs, bounds, or derivations? **7.** Does the report connect claimed dataset or benchmark contributions to concrete properties such as scale, diversity, annotation process, coverage, or evaluation method? **8.** Does the report mention important objective functions, losses, constraints, assumptions, equations, or proof targets when they are central to the paper? **9.** Does the report identify the datasets, benchmarks, domains, or experimental environments used by the main paper rather than describing evaluation generically? **10.** In the 'Similarities' section, does the report describe specific shared mechanisms, equations, or structural elements between the main paper and the retrieved works, rather than just generic thematic overlaps? **11.** If specific baselines, prior models, datasets, systems, or theoretical frameworks are mentioned, does the report compare the current paper against the strongest or closest ones rather than only convenient examples? **12.** Does the comparative analysis distinguish substantive innovation from changes in terminology, scale, benchmark choice, implementation detail, or packaging? **13.** When a cited related paper is described as different, does the report identify the exact missing, changed, or weaker component in that prior work that supports the contrast? **14.** When multiple related papers are cited for one novelty point, does the report distinguish the role of each cited paper in the comparison rather than treating them as interchangeable background? **15.** For the closest prior works, does the report discuss them in enough technical detail to justify whether the current contribution is substantially different or mainly incremental? **16.** When making an originality judgment, does the report connect that judgment to concrete overlaps or extensions relative to cited prior work rather than relying on broad field-level statements? **17.** When a related paper is cited for a dataset or benchmark, does the report identify the relevant dataset property, task, scale, or evaluation setting? **18.** When a related paper is cited for a theoretical result, does the report identify the relevant theorem, assumption, framework, or analytical conclusion? **19.** When a related paper is cited for an empirical finding, does the report identify the relevant observation, trend, metric, or experimental setup? **20.** In the Similarities section, does the report identify shared system components, pipeline stages, or infrastructure choices when relevant? **21.** Does the report make clear which cited paper supports each major difference claim? **22.** Does the report explain why the similarity matters for evaluating novelty rather than merely noting that two works are related? **23.** When the provided text highlights similarities between the main paper and prior work, does the report's final innovation assessment reflect these similarities, rather than simply repeating the author's claims of novelty? **24.** Does the report evaluate innovation along multiple dimensions–such as implementation, similarity to prior work, and technical contribution–based on the evidence in the provided texts? **25.** Does the report appear to autonomously assess the author's claims on the basis of evidence from the provided texts, rather than just repeating the author's words? **26.** Does the innovative summary acknowledge at least one limitation, boundary condition, or negative aspect of the claimed innovations, if such information is inferable from the provided texts? **27.** Does the report identify standard practice or routine engineering choices when the comparison evidence supports that judgment?

Figure 13: Depth checklist questions used in checklist evaluation.

Checklist Evaluation Answering Prompt

Task overview. The evaluator is instructed to read a provided report and a database containing information on related papers, and then answer `yes` or `no` to specific checklist questions. The questions correspond to a particular evaluation dimension of the report.
Dimension specification. The prompt provides the target dimension in the form: `{dimension} - {definition}`, followed by dimension-specific conditions, denoted as `{conditions}`.
Instructions.

- Read the instructions thoroughly.
- Carefully read both the report and the database.
- Understand the checklist questions and the definition of the target dimension.
- Respond to each question with `yes` or `no`.
- Base each answer on a clear rationale.
- Follow the specified answer format.

Evaluation focus note. The prompt clarifies that the report may contain different kinds of content, including statements that summarize the main paper and statements that compare it with retrieved or cited related works. The exact naming of a group is less important than its evaluation focus. The evaluator should use the provided `{focus_note}` for the current question group.
Equivalent naming note. When evaluating structure or section-specific questions, the evaluator should judge flexibly according to the actual function and meaning of a section or subsection. If two titles express nearly the same function but use different wording, they should be treated as equivalent rather than marked as missing merely because of wording differences. In particular:

- *Comparative Analysis*, *Point-wise Novelty Analysis*, and *Point-wise Comparative Analysis* are equivalent titles for the second main section.
- *Innovation Point* and *Novelty Point* are equivalent titles for each point-level subsection.
- *Claimed Innovation* and *Claimed Novelty* are equivalent subsection titles.
- *Details of Unique Difference*, *Details of Unique Differences*, and *Unique Feature Description* are equivalent subsection titles.
- *Innovation Summary* and *Novelty Summary* are equivalent titles for the final main summary section.
- *Genuine Innovations* and *Genuine Novelties* are equivalent subsection titles.
- *Aspects Lacking Originality*, *Aspects Lacking Substantial Originality*, and *Aspects lacking substantial originality* are equivalent subsection titles.
- *References* and markdown reference-link definitions are acceptable citation-support sections and should not be treated as extra analytical sections.

Answer format. The evaluator must return one line per checklist question using the required format: `Q1: [Your Answer]`, `Q2: [Your Answer]`, and so on, where each answer is `yes` or `no`.

Figure 14: Core prompt used by the evaluator to answer checklist questions.

Point-wise Novelty Analysis Prompt

Please perform a detailed point-wise novelty analysis for the following novelty point extracted from the paper `{paper_name}`. `Novelty Point {point_num}: {innovation_point}` **INSTRUCTIONS:**

- Use the Novelty Point description to understand the current claim's content, methodology, evidence, and claimed contribution.
- Use the retrieved prior-work evidence to compare against concrete prior papers and to support cited claims.
- Use the historical prior-work context from the system message as historical comparison evidence. Do not name or describe this source in your output, but when you rely on a concrete historical paper's innovation content or evaluation, cite its provided Historical citation marker.
- Use the selected leaf-theme histories for detailed paper-level comparison, and use the broader parent-theme overall summary only to understand the larger development trajectory of the field.
- In Similarities and Unique Differences, do not only say that the current point is in the same broad theme. Explain the content-level overlap or difference: what mechanism, dataset, theoretical claim, empirical finding, task formulation, or system capability is shared or different.
- If the historical context suggests that the current point is largely a small variation of prior historical innovations, say this as a natural judgment about prior work, not as a statement about the context source.
- Keep the output structure stable. Output each of the four subsections exactly once. Do not repeat b/c/d sections. Do not add extra Section 2 headings inside your response.
- Do not include meta-commentary, correction notes, validation notes, or source-debugging notes.
- Forbidden output phrases include: "expert knowledge", "expert memory", "memory system", "RAGFlow", "system message", "system context", "retrieved related texts", "provided materials", "corrected for", and "The original reference".

OUTPUT FORMAT: a) Restatement:
Rephrase the novelty point clearly and in detail based on the novelty description. Include methodology, architecture, dataset/benchmark, theoretical claim, empirical evidence, or system details when they are available. **b) Similarities:**
Compare this novelty point with prior work. If no similarities are found, state: "Based on the available prior-work evidence, no explicit similarities with existing work were identified." Otherwise, describe specific shared objectives, mechanisms, datasets, evidence types, or problem formulations. **c) Unique Differences:**
Excluding the similarities in part (b), analyze what makes this novelty point different from prior work. Use specific details from the novelty point description. If no unique differences are found, state: "Based on the available prior-work evidence, no unique differences were identified for this novelty point." **d) Details of Unique Differences:**
If genuinely unique aspects were identified in part (c), analyze their significance in one paragraph. Follow the innovation type: for methods, explain the structural/procedural difference and claimed benefit; for theory, explain the central claim and rigor; for systems, explain the solved problem and demonstrated practicality; for datasets/benchmarks, explain the resource gap and enabled capabilities; for empirical/analytical points, explain the central finding and experimental support; for tasks/applications, explain why the task/application is important and non-obvious. If part (c) found no uniqueness, leave only a brief sentence explaining that no distinct details can be added. **Important Reminder:** During Similarities, Unique Differences, and Details, compare the current novelty point against prior work. Do NOT compare the novelty point with itself. Do NOT fabricate implementation details, metrics, datasets, or claims not present in the available evidence. Now generate the analysis.

Figure 15: Prompt used for point-wise novelty analysis against prior work.

Prompt used to generate the novelty summary

You are an Expert Novelty Assessment Specialist with deep expertise in evaluating the novelty, significance, and impact of research contributions. Your role is to synthesize point-wise comparative analyses into a balanced final novelty summary. You will receive two inputs:
1. Section 2: point-wise novelty analysis for each claimed innovation.
2. Historical prior-work context corresponding to the innovation points.

How to use the historical context:
- Use it to understand how similar historical innovations were evaluated, including their ratings, assessment text, and historical standing.
- Use these historical evaluations as reference examples for calibrating whether the current paper is transformative, significant, incremental, or weak.
- Distinguish between claimed innovation points that have little substantial originality and claimed innovation points that are genuinely innovative.
- ...
- Do not name this context source or expose internal pipeline mechanics.

Important restrictions:
- DO NOT use any citations or references in Section 3.
- Base your analysis only on Section 2 and the historical context below.
- Use general terms like "existing work", "prior methods", "historical innovations", and "related approaches".
- ...

Section 2 Content:
`{draft_section2}` **Historical Context:**
`{expert_knowledge_bundle}` **Generate Section 3 exactly as follows:**
`## 3. Novelty Summary`
Write a comprehensive, detailed, and faithful synthesis of the paper's novel contributions. First articulate the paper's overall innovative characteristics. Then distinguish genuinely innovative contributions from aspects that lack substantial originality. **Required content:**
- **Genuine Innovations:** Identify the innovation point(s) that are genuinely innovative, explain the core unique idea, and discuss its significance in historical context.
- **Aspects Lacking Originality:** Identify claimed innovations that are not substantially original. Categorize each as Direct Repetition, Minor Adjustment, Simple Combination, Standard Practice, or Other.

Final One-line Summary:
End with exactly one final one-line summary in this format: `**<score> - <label>:** <one sentence explaining the overall novelty, strengths, and limitations>`. The score should be calibrated using both Section 2 and the historical evaluation examples. **Rating rules:**
- **4 - Excellent / Transformative:** Entirely new question, highly novel methodology, or surprising perspective.
- **3 - Good / Significant:** Clear and substantial progress through a non-obvious methodology or creative combination.
- **2 - Fair / Incremental:** Limited extension through minor improvements, tweaks, or known techniques.
- **1 - Poor / Insufficient:** Lacks clear original contributions or presents well-known, trivial, or previously published ideas.

Strict output rules:
- Output ONLY Section 3 content, starting with "## 3. Novelty Summary".
- The final line must include the score, label, and one explanatory sentence.
- Be concise, neutral, and faithful to Section 2 and the historical context.
- DO NOT use citations or paper-reference markers.
- ...

Figure 16: Prompt used to generate the novelty summary.

Memory Development-History Prompt

System prompt. You are an expert in structuring the development history of academic-paper innovation points. You are skilled at integrating publication time, paper titles, summaries, innovation claims, final decisions, and peer-review signals into a precise, high-detail, structured expert record. Your output will later be used for visualization, analysis, and as expert knowledge for another LLM to judge the novelty of a new innovation point and compare it against prior work. Therefore, you must preserve concrete technical content, accurately describe what each innovation point actually is, make the evolution of the category clear over time, and output a simple JSON object. You must cover every innovation point and must not omit any point. You should use review information selectively and critically: prioritize PC/AC meta-review over conflicting reviewer comments, use reviewer scores together with comment language and final decision to judge historical standing, ignore review comments that are irrelevant to the specific innovation point being discussed, and reason independently when the review evidence is absent, weak, or clearly unrelated. **User prompt.** You are given one category of academic innovation points, together with their source documents, publication years, paper titles, paper summaries, final decisions, and review comments. **Category metadata:**

- Theme ID: `{theme_id}`
- Theme Title: `{theme_title}`
- Theme Description: `{theme_description}`

Your task:

1. Output ONLY a JSON object. Do not output markdown, explanation, or prose outside JSON.
2. The JSON object must follow this schema exactly:

Required JSON schema: `{theme_overview; innovation_points: [{item_number, year, paper_title, innovation_content, evaluation: {rating, assessment}}]; overall_evolution_summary}`.

3. The `innovation_points` list must be in chronological order as much as possible.
4. For EACH innovation point, explain its concrete content in enough detail for downstream comparison, following the original innovation-point text closely, and assess its historical standing in the line of work.
5. Use the paper summary, paper title, and innovation-point text as the primary evidence; use review comments, reviewer scores, meta-review, and final decision as auxiliary evidence for historical standing and significance, not as blind substitutes for reasoning.
6. ...

Important constraints:

- Only use the provided source material.
- Preserve chronological order as much as possible.
- The result should be structured JSON, not prose outside JSON.
- Do not include extra keys beyond the schema above.
- The history should be detailed enough to support downstream novelty judgment, side-by-side comparison, timeline visualization, and expert-memory use.
- Do not say that some points were omitted, merged, or unavailable.
- When using review signals, do not quote them mechanically without interpretation; integrate them into your own expert judgment.

Source material:

`{source_material_text}`

Figure 17: Prompt used to write leaf-level development histories in the historical research memory.