Agentic Paper Reviewer: Academic Paper Reviewer Based on AI Agents

Suho Han [1]

Abstract

Peer review provides high-quality feedback but suffers from slow iteration cycles and high variance in review usefulness. We present REVIEWERAGENT, an on-premise, retrieval-augmented, multi-agent paper reviewing system that connects public OpenReview review corpora to a LangGraph-orchestrated agent pipeline. The system combines (i) structured retrieval over historical human reviews using a ChromaDB vector store and (ii) coordinated reviewer and rating agents to produce actionable feedback and rubric-style assessments. To make experimentation practical, we self-host lightweight (1B) and stronger (8B) open-source instruction-tuned LLMs via vLLM and provide a Streamlit demo. We evaluate against a zero-retrieval baseline on a held-out test set and report metrics including weakness recall, score correlation, and hallucination rate.

1. Introduction

The peer-review process is the primary mechanism for validating and improving research papers, yet it is an unusually slow feedback loop: authors may wait months between submission and receiving comments, and the resulting feedback can be noisy or non-actionable. This is especially painful for early-stage projects where rapid iteration on experiments, writing clarity, and positioning relative to prior work is critical.

Large language models (LLMs) have recently become strong assistants for writing and analysis, but naive "single-shot" review generation can be brittle: it may miss key weaknesses, overlook relevant baselines, or fabricate supporting evidence. We hypothesize that a useful reviewer assistant should (i) ground feedback in retrieved evidence and (ii) follow an explicit multi-step workflow that mirrors how human reviewers read, cross-check, and synthesize.

We introduce REVIEWERAGENT, a retrieval-augmented, LangGraph-based multi-agent pipeline for paper review generation from OpenReview-derived data.

REVIEWERAGENT is designed to run fully on-premise with self-hosted vLLM endpoints, enabling reproducible experiments without reliance on proprietary APIs.

Our key contributions are:

- **An on-premise agentic reviewing pipeline** that integrates RAG over OpenReview human review corpora with a coordinator-driven multi-agent workflow (retrieval, review drafting, and rating).
- **A practical, reproducible system implementation** with ChromaDB vector stores, vLLM-hosted open-source LLMs (1B/8B), and an end-to-end Streamlit demo.
- **An evaluation protocol for reviewer agents** including a baseline-vs-agent comparison and metrics that target review usefulness (weakness recall), faithfulness (hallucination rate), and calibration (score correlation).

2. Related Work

We position this work at the intersection of retrieval-augmented generation (RAG), agentic AI (tool-using and multi-step LLM systems), and emerging "agentic" workflows for assisting scientific reviewing.

Retrieval-Augmented Generation (RAG). RAG combines parametric generation with non-parametric retrieval to improve factuality and coverage on knowledge-intensive tasks (Lewis et al., 2020). Subsequent work explores stronger fusion architectures and scaling retrieval+generation pipelines, including fusion-in-decoder style conditioning (Izacard & Grave, 2021) and other retrieval-conditioned generators. In the context of paper understanding and review assistance, RAG is particularly relevant because it enables grounding claims in retrieved evidence (e.g., paper sections, cited works, or external corpora), which can reduce hallucinations and support verifiable summaries.

Agentic AI and tool-using LLM systems. Agentic AI systems extend LLMs beyond single-shot generation by enabling iterative reasoning, tool use, and multi-step planning/execution (Yao et al., 2023; Schick et al., 2023; Wu et al., 2023). A representative line of work uses interleaved

reasoning and acting (e.g., calling search, code, or structured tools) to solve complex tasks (Yao et al., 2023). These agentic paradigms are a natural fit for reviewing, where an assistant may need to (i) retrieve evidence, (ii) verify claims, (iii) cross-check prior work, and (iv) synthesize structured feedback under constraints.

Agentic reviewer workflows (practitioner systems). Parallel to academic research, practitioner-facing systems have proposed structured "agentic" workflows for document analysis and reviewing. For example, the Stanford "Agentic Reviewer" system (PaperReview.ai) converts a paper PDF into a structured representation, retrieves and summarizes relevant prior work from arXiv via web search, and then generates a comprehensive review following a template; it also reports rubric-style sub-scores and studies agreement with public ICLR 2025 ratings (Jiang & Ng, 2025). We treat these workflows as complementary motivation: they demonstrate practical decomposition patterns (extraction $\rightarrow$ retrieval $\rightarrow$ synthesis $\rightarrow$ scoring) that can be operationalized and evaluated in a research setting.

3. Method

3.1. Proposed Approach

REVIEWERAGENT is implemented as a LangGraph workflow composed of specialized agents coordinated by a top-level controller. At a high level, the pipeline proceeds as follows:

1. **Ingestion and indexing:** We collect OpenReview review data (e.g., ICLR/NeurIPS/ICML venues) and build a ChromaDB vector store over review text.
2. **Retrieval (RAG):** Given a query derived from the target paper (e.g., "main weaknesses in experimental design"), the retriever agent fetches semantically similar historical reviews and optionally augments with external retrieval (e.g., arXiv search) when enabled.
3. **Review generation:** A reviewer agent synthesizes a structured review grounded in the retrieved evidence.
4. **Rating:** A rating agent produces rubric-style assessments and an overall score, aiming to be consistent with the evidence and the review text.
5. **Coordination and iteration:** A coordinator agent can request additional retrieval or refinement when missing information is detected.

3.2. Training and Implementation Details

The system uses instruction-tuned open-source LLMs served via vLLM. In our default configuration, we run a lightweight model (e.g., Llama-3.2–1B–Instruct) and a stronger model (e.g., Llama-3–8B–Instruct) on separate ports with configurable GPU memory utilization and maximum context length. All components are packaged as scripts for data collection, vector DB building, and evaluation, and a Streamlit UI for interactive demos.

4. Results

Baselines. We compare REVIEWERAGENT against a **zero-retrieval baseline** that disables RAG and external search, using only the reviewer and rating components.

Metrics. We report (i) **weakness recall** against a reference set of weaknesses, (ii) **rating correlation** (Pearson/Spearman) between predicted and human reference scores, and (iii) **hallucination** metrics computed from an automated claim-checking procedure: a scalar hallucination score and the hallucinated-claim ratio.

Protocol. We run both baseline and agent pipelines on the same sampled subset and aggregate results into a JSON report and a markdown summary.

External comparison (reported results). To contextualize our correlation results, we additionally report headline numbers from the PaperReview.ai "Agentic Reviewer" technical overview (Jiang & Ng, 2025). Their system uses an agentic workflow that converts PDFs to markdown, retrieves prior work from arXiv via web search, and produces a review plus a 7-dimension rubric (e.g., originality, soundness of experiments, clarity, and contextualization) which is then mapped to an overall score using linear regression. On a random sample of ICLR 2025 submissions, they report that the AI score achieves Spearman correlation comparable to human–human agreement and provides non-trivial acceptance prediction AUC (Table 2). We emphasize these numbers are *not directly comparable* to ours due to different datasets, access to external web search/arXiv grounding, and differing score construction protocols.

4.1. Main Results

Table 1 summarizes the baseline-vs-agent comparison on $N = 899$ examples. Overall, enabling retrieval and agentic coordination yields a small increase in weakness recall ($0.0084 \rightarrow 0.0096$). However, in this run, hallucination increases substantially (hallucination ratio $0.0636 \rightarrow 0.1555$) and rating correlation becomes more negative (Spearman $-0.059 \rightarrow -0.427$). These results suggest that while retrieval can help surface some missing weaknesses, additional constraints (e.g., stricter grounding/citation requirements and calibration of the rating head) are required to improve faithfulness and score alignment.

Table 1. Baseline vs. REVIEWERAGENT on the OpenReview-derived test set ($N = 899$). Correlations are computed between predicted and mean human ratings. Lower hallucination is better; higher weakness recall is better.

Metric	Baseline	Agent
Pearson (r)	-0.0786	-0.4307
Spearman (ρ)	-0.0593	-0.4271
Avg. weakness recall	0.0084	0.0096
Avg. hallucination score	0.0895	0.2098
Avg. hallucination ratio	0.0636	0.1555

Table 2. PaperReview.ai reported score-agreement metrics on ICLR 2025 submissions (randomly sampled 300 submissions; linear-regression score mapping trained on 150 and tested on 147; from (Jiang & Ng, 2025)). These results are shown only for context and are not directly comparable to Table 1.

Metric	Human vs. human	AI vs. human
Spearman (ρ) score correlation	0.41	0.42
AUC for predicting acceptance	0.84	0.75

4.2. Ablation Studies

We plan ablations that isolate the effect of (i) RAG over OpenReview reviews, (ii) optional external retrieval, and (iii) model size (1B vs. 8B). We leave these ablations to future work as they are not included in the current evaluation run.

4.3. Analysis

We observe three common failure modes: (i) retrieval misses when the paper is out-of-domain relative to indexed venues, (ii) overly generic feedback when the manuscript lacks concrete experimental detail, and (iii) brittle scoring when the model conflates novelty with writing quality. On the systems side, on-prem vLLM hosting enables predictable latency and cost, but requires careful GPU memory and context-length configuration.

5. Conclusion

We presented REVIEWERAGENT, an on-premise, retrieval-augmented multi-agent pipeline that connects OpenReview human review corpora to a LangGraph workflow for generating structured, actionable paper reviews. The system is practical to run with self-hosted vLLM models and includes scripts for dataset construction, vector database building, and baseline-vs-agent evaluation.

6. Related Work

How this work fits. Relative to standard RAG pipelines (Lewis et al., 2020), our focus is not only on grounded generation but also on end-to-end reviewing actions (e.g., targeted retrieval, claim checking, and structured decision support). Relative to general agent frameworks (Wu et al., 2023; Yao et al., 2023), we specialize the agent loop and evaluation to the peer-review context.

References

Izacard, G. and Grave, E. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

Jiang, Y. and Ng, A. Tech overview: Stanford agentic reviewer (paperreview.ai). https://paperreview.ai/tech-overview, 2025. Accessed: 2025-12-18.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K"uttler, H., Lewis, M., Yih, W., Rockt"aschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. 2023. doi: 10.48550/arXiv.2302.04761. URL https://arxiv.org/abs/2302.04761.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. 2023. doi: 10.48550/arXiv.2308.08155. URL https://arxiv.org/abs/2308.08155.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.