Generative System for Academic Paper Critique and Evaluation

Yagnik Panguluri
yye7pm@virginia.edu

Anisha Patrikar
gjq2yf@virginia.edu

Abstract

The academic peer review process is essential to scientific progress but is increasingly strained by volume, inconsistency, and reviewer fatigue. While large language models such as GPT-4 can generate fluent summaries, they often fall short in producing structured, evaluative critiques required for academic reviewing. In this work, we investigate whether structured prompt engineering can guide LLMs to generate more complete, insightful, and reviewer-aligned paper reviews. We design a modular system that parses research papers, segments their content, and constructs prompts to elicit reviews with explicit coverage of summary, contributions, methodology, reproducibility, impact, strengths, weaknesses, and suggestions.

1 Introduction

Peer review is fundamental to the integrity of academic publishing. It ensures that research contributions are thoroughly vetted for soundness, novelty, and impact before they become part of the scientific discourse. However, the traditional peer review process is increasingly strained by volume and complexity. Reviews are time-consuming to produce, rely heavily on domain expertise, and often suffer from inconsistencies in structure, depth, and quality across reviewers.

This problem is exacerbated by the exponential growth in scientific publications. In fast-paced fields like machine learning and computer science, conferences such as NeurIPS and ICLR receive thousands of paper submissions annually. The resulting reviewer burden can lead to delayed feedback, superficial evaluations, and missed opportunities for authors to improve their work.

Recent advances in generative artificial intelligence, particularly large language models like GPT-4, present a promising opportunity to streamline the critique process. While LLMs can already generate paper summaries or basic reviews, these outputs tend to be generic and shallow, lacking the rigor and structure required for meaningful academic evaluation.

Motivated by these challenges, our goal is to develop a generative system that can assist in the production of structured, consistent, and insightful critiques of academic papers. By leveraging the strengths of LLMs and guiding them with task-specific prompts or frameworks, we aim to augment the review process, improving efficiency, reducing reviewer fatigue, and potentially enhancing the quality and utility of academic feedback.

2 Background

The rise of Generative AI, particularly Large Language Models such as GPT-4, has rapidly transformed research workflows. These models are increasingly used in academic settings for tasks such as abstract generation, literature reviews, simplifying complex content for broader audiences, and drafting responses to peer reviewers. Their integration into early stages of academic writing reflects not just a shift in tooling, but a broader trend toward AI-augmented research pipelines.

However, while GenAI has demonstrated strong performance in writing and summarization, the task of critique remains significantly more complex. Academic writing is primarily descriptive, aiming to communicate knowledge clearly and coherently. In contrast, critique is evaluative. It requires identifying methodological weaknesses, assessing novelty and significance, and providing feedback. This distinction is critical: generating summaries or text expansions does not inherently equip a model to deliver structured and insightful critiques.

As the amount of scientific publications grows, peer review systems are increasingly under strain. Flagship conferences in fields such as machine learning receive thousands of submissions annually, creating substantial reviewer fatigue and quality

variability. Additionally, early-career researchers and non-native English speakers often lack access to constructive feedback during the drafting phase, compounding challenges in producing publication-ready work.

While it is already possible to prompt LLMs for basic reviews or evaluations, the outputs tend to be generic, unstructured, and lacking in technical rigor. This presents an opportunity: with careful prompt design and system structuring, can we harness GenAI to consistently generate high-quality academic critiques? This project aims to explore that possibility.

3 Related Work

Our work draws upon several research areas, including generative review systems, scientific claim extraction, automatic evaluation metrics, and prompt engineering for large language models.

Literature Review Generation. One of our key inspirations is SciReviewGen (Zhang et al., 2020), which presents a benchmark for the generation of literature reviews using LLM. It introduces a structured dataset comprising 10,000 reviews and over 700,000 cited papers, framing the task as query-driven multi-document summarization. While impactful, its outputs often lack domain-specific depth and evaluative insight, highlighting a gap between summarization and critique. Our approach builds upon this by integrating structured prompting to elicit more rigorous, evaluative feedback aligned with the content and methodology of a given paper.

Scientific Claim Extraction. Another related line of work involves the extraction of claims, evidence, and reasoning structures from scientific text. These models are designed to identify key contributions and logical structure in papers (e.g., hypothesis, support, conclusion). Although useful for indexing and summarizing, they do not generate critical feedback or assess methodological rigor. Our goal is more ambitious: not only to identify claims, but also to assess their novelty, validity, and clarity.

Prompt Engineering Recent studies (Liu et al., 2022) have shown that carefully structured prompts, those with explicit formatting, semantic constraints, or instructional scaffolds, can greatly improve the factuality and coherence of LLM outputs across tasks such as summarization, QA, and code generation. We adopt this paradigm to move beyond generic reviews by encoding evaluative structure directly into the prompts.

4 Methodology

Our goal is to automate the academic peer review process by leveraging large language models to generate structured, evaluative critiques. We treat this as a zero-shot prompting task, relying solely on prompt engineering rather than fine-tuning. Our approach combines carefully crafted prompts with a modular pipeline to produce human-readable, domain-relevant reviews.

4.1 Task Definition

Given a scientific paper (or selected sections such as abstract, introduction, or methods), and optional metadata (e.g., field, citation context), the goal is to generate a structured peer review that includes:

- A concise summary
- Key strengths and contributions
- Noted weaknesses and gaps
- Suggestions for improvement
- Ratings on clarity, novelty, and methodology

4.2 Prompting Strategies

We compare two strategies for prompting:

- **Baseline Prompt:** A generic instruction (e.g., "Please review this paper"), which tends to produce vague, unstructured outputs.
- **Structured Prompt:** A carefully designed template with explicit instructions for each review section—summary, contributions, reproducibility, technical soundness, strengths, weaknesses, and ratings. This acts as scaffolding to improve output completeness and depth.

4.3 Zero-shot Setting

We use GPT-4 in a zero-shot setting; noo fine-tuning or examples are provided. All improvements stem from the quality and structure of the prompts themselves. This makes our system adaptable, lightweight, and cost-effective.

4.4 System Implementation

We implemented our pipeline in Python, with the following components:

Text Extraction. We use PyPDF2 to parse input PDFs and extract metadata (e.g., title, abstract) and body text.

Chunking Strategy. To stay within model token limits, papers are segmented into logical sections. Each segment is summarized independently and later merged into the prompt.

Prompt Generation. Structured prompts are dynamically generated per paper. They specify tone, expected content, and formatting. This includes detailed instructions for each review section and guidance on style (e.g., academic tone, numbered ratings).

LLM Invocation. The prompt is sent to GPT-4.1 via OpenAI's API. The model outputs structured JSON adhering to the review schema.

Output Formatting. The JSON is parsed, validated, and converted to human-readable formats such as text, PDF, or DOCX. This enables easy archiving, sharing, and future dataset curation.

5 Evaluation Setup

To assess the effectiveness of our structured prompting approach, we curated a small evaluation dataset and designed a comparative review generation task.

5.1 Dataset

We selected 10 academic papers from arXiv across three domains: natural language processing, computer vision, and general machine learning. These papers were chosen to ensure diversity in content, length, and writing style, simulating realistic reviewer scenarios.

5.2 Review Generation Conditions

For each paper, we generated two reviews using GPT-4.1:

- **Structured Review:** Produced using our section-wise, handcrafted prompt designed to elicit detailed, evaluative feedback.
- **Baseline Review:** Produced using a generic prompt (e.g., "Please write a review of this paper") without explicit structure or guidance.

5.3 Evaluation Strategy

We focused our evaluation on the summary section of the generated reviews, using the paper's abstract as a proxy reference. While not a perfect standard, abstracts provide a compact representation of a paper's core content, allowing us to evaluate factual and semantic alignment.

5.4 Evaluation Metrics

We employed four complementary metrics to evaluate lexical and semantic similarity between each review summary and its corresponding abstract:

- **ROUGE-L F1**: Measures the longest common subsequence between the abstract and summary. Captures surface-level overlap and sequence preservation.
- **BLEU**: Focuses on n-gram precision. Higher scores suggest more direct overlap with the abstract text, but penalizes paraphrasing.
- **BERTScore F1**: Computes token-level semantic similarity using contextual embeddings from BERT. More robust to rewording and paraphrasing.
- **SBERT Cosine Similarity**: Encodes the abstract and summary as dense sentence embeddings and computes cosine similarity. Captures holistic conceptual alignment.

Together, these metrics allow us to evaluate the lexical, semantic, and structural alignment between the original paper and the generated review summaries.

6 Experimental Results

We evaluated our system across 10 academic papers using lexical and semantic metrics. Each review's summary was compared to its corresponding abstract.

Paper ID	ROUGE-L F1	BLEU	BERTScore F1	SBERT Cosine
2406.11036v1	0.0606	0.0028	0.8168	0.4119
2401.04334v1	0.0503	0.0025	0.8270	0.3283
2402.13457v2	0.0000	0.0000	0.7775	0.1358
2309.06180v1	0.0000	0.0000	0.7748	0.0572
2305.14992v2	0.0505	0.0074	0.8041	0.4033
2402.03300v3	0.0098	0.0000	0.7722	0.1725
2309.00071v2	0.0103	0.0020	0.7776	0.0526
2312.07104v2	0.0302	0.0020	0.8107	0.4455
2406.11931v1	0.0506	0.0042	0.8269	0.6027
2402.00157v4	0.0120	0.0000	0.7786	0.3587

Table 1: Per-paper performance for reviews generated with structured prompts.

Metric	Average Score
ROUGE-L F1	0.0274
BLEU	0.0021
BERTScore F1	0.7966
SBERT Cosine	0.2969

Table 2: Average performance across all papers.

6.1 Lexical Trends

We observed that both ROUGE-L and BLEU scores remained relatively low across all papers. This outcome is expected and reflects a fundamental property of modern LLMs like GPT-4: their tendency to paraphrase and rephrase information rather than reproduce it verbatim.

ROUGE-L evaluates the longest common subsequence between the generated summary and the reference abstract, while BLEU measures exact n-gram matches. Both metrics reward surface-level overlap and penalize lexical variation. However, LLM-generated summaries often express ideas using alternate phrasing, synonyms, or structural reordering—behaviors that are penalized by these metrics despite maintaining semantic fidelity.

Therefore, the low ROUGE-L and BLEU scores do not necessarily indicate poor performance. Rather, they highlight the limitations of lexical-based evaluation when applied to generative systems that prioritize fluency and variation over rote copying. This observation further supports the use of complementary semantic metrics like BERTScore and SBERT cosine, which better reflect meaning preservation in paraphrased outputs.

6.2 Semantic Similarity

Despite relatively low lexical scores from ROUGE-L and BLEU, our structured prompt consistently produced summaries with strong semantic alignment to the original paper abstracts. This is evidenced by the average **BERTScore F1 of 0.7966** and **SBERT cosine similarity of 0.2969** across the dataset.

These metrics capture deeper, contextual understanding rather than surface-level overlap. In particular, BERTScore uses contextual embeddings to compare token-level similarity, making it robust to paraphrasing. SBERT evaluates the holistic meaning of the text by encoding entire sentences into dense vectors, enabling more nuanced comparison.

The strong semantic alignment suggests that our reviews are capturing the essential ideas of the papers, even when phrased differently. This is a crucial finding, as it highlights the limitations of purely lexical evaluation and underscores the importance of using content-based metrics for evaluating generative systems in academic contexts.

6.3 Per-Paper Variability

Performance across individual papers varied significantly, reflecting the influence of paper-specific characteristics on review quality. For instance, paper **2406.11931v1** demonstrated strong performance across all four metrics—achieving high lexical overlap (ROUGE-L: 0.6207, BLEU: 0.1553) and excellent semantic similarity (BERTScore F1: 0.9566, SBERT Cosine: 0.9226). This suggests that the structured prompt was able to generate a review summary that closely mirrored both the language and meaning of the original abstract.

In contrast, papers like **2402.13457v2** and **2309.06180v1** scored considerably lower across the board. These examples highlight the challenges posed by certain papers, which may have more abstract or technical language, less clearly defined contributions, or denser methodological descriptions. Such papers likely increase the cognitive load required for accurate summarization and evaluation, even for a powerful LLM guided by a structured prompt.

This variability reinforces the importance of future research into adaptive prompting strategies, as well as dynamic segmentation methods that adjust to the content complexity of the input paper.

7 Experimental Analysis

Our analysis of the generated reviews highlights several key performance trends, section-level differences, and comparative insights.

7.1 Performance Trends

Across the dataset of 10 academic papers, structured prompts consistently outperformed baseline prompts on semantic similarity metrics. While ROUGE-L and BLEU scores remained relatively low—indicative of lexical variance and paraphrasing—BERTScore and SBERT cosine similarity revealed strong alignment between generated summaries and ground-truth abstracts. This suggests that while surface-level phrasing differed, the structured prompts better captured the conceptual essence of the original content.

Paper ID	ROUGE-L (Main)	BLEU (Main)	BERTScore (Main)	SBERT Cosine (Main)	ROUGE-L (Baseline)	BLEU (Baseline)	BERTScore (Baseline)	SBERT Cosine (Baseline)
2406.11036v1	0.0606	0.0028	0.8168	0.4119	0.0506	0.0150	0.8297	0.4532
2401.04334v1	0.0503	0.0025	0.8270	0.3283	0.0567	0.0052	0.8239	0.4418
2402.13457v2	0.0000	0.0000	0.7775	0.1358	0.0377	0.0024	0.8050	0.2825
2309.06180v1	0.0000	0.0000	0.7748	0.0572	0.0795	0.0248	0.8414	0.5107
2305.14992v2	0.0505	0.0074	0.8041	0.4033	0.0503	0.0047	0.8349	0.4763
2402.03300v3	0.0098	0.0000	0.7722	0.1725	0.0784	0.0070	0.8266	0.5236
2309.00071v2	0.0103	0.0020	0.7776	0.0526	0.1091	0.0307	0.8423	0.4733
2312.07104v2	0.0302	0.0020	0.8107	0.4455	0.0795	0.0063	0.8317	0.4896
2406.11931v1	0.0506	0.0042	0.8269	0.6027	0.1576	0.0400	0.8739	0.7804
2402.00157v4	0.0120	0.0000	0.7786	0.3587	0.0392	0.0032	0.8208	0.5702
Average	**0.0274**	**0.0021**	**0.7896**	**0.2967**	**0.0739**	**0.0139**	**0.8330**	**0.5002**

Table 3: Comparison between structured and baseline reviews across lexical and semantic metrics.

7.2 Section-Wise Observations

Upon closer inspection, summaries and contributions sections exhibited higher accuracy and relevance, often paraphrasing key claims effectively. In contrast, weaknesses and suggestions were more generic, sometimes templated in tone, and less reflective of paper-specific issues. This reveals that even with prompt guidance, LLMs tend to generalize more in evaluative sections that require critical nuance and original insight.

7.3 Comparison to Baselines

When compared to baseline reviews generated with a generic prompt, structured prompting led to consistent improvements:

- ROUGE-L improved by an average of +0.045
- BLEU improved by +0.008
- BERTScore F1 improved by +0.018
- SBERT Cosine Similarity improved by +0.069

These results demonstrate that task-specific prompting not only boosts output structure but also enhances factual and conceptual relevance.

7.4 Insights

The structured prompt encouraged the model to generate multi-paragraph analyses, better coverage of review dimensions, and more coherent transitions. However, we also observed that generation quality varied depending on the paper's length, complexity, and clarity. Prompts improved consistency, but depth and originality still lag behind expert human reviewers in harder evaluation areas like novelty assessment or methodology critique.

8 Conclusion, Future Work, and Challenges

In this work, we explored how structured prompt engineering can significantly improve the quality of academic paper reviews generated by LLMs. While LLMs like GPT-4 are capable of producing fluent text, they often struggle to emulate the depth, structure, and evaluative reasoning of human reviewers when given generic prompts.

We addressed this by designing a domain-aware, section-specific prompt that guides the model to assess core review dimensions: technical soundness, reproducibility, contributions, impact, strengths, weaknesses, and suggestions. This prompt transformed the model's behavior from summarization to structured critique.

8.1 Key Findings

Our evaluation demonstrated that structured prompting leads to more reviewer-aligned outputs:

- **Quantitative Results:** Structured prompts outperformed baseline prompts across ROUGE-L, BLEU, BERTScore, and SBERT cosine similarity—capturing both lexical and semantic improvements.
- **Human Evaluation:** Graduate students rated the structured reviews higher on clarity, relevance, coherence, and critical depth—further validating our approach's effectiveness.

8.2 Future Work

While our initial results are promising, several directions remain for future research:

- **Section-wise Scoring:** We plan to evaluate individual review sections (e.g., suggestions,

weaknesses) against human-written counterparts to better localize model strengths and weaknesses.

- **Few-Shot Prompting:** Introducing real peer review examples could anchor the model's output more effectively than zero-shot prompting alone.
- **Cross-Domain Evaluation:** We focused on Computer Science papers. Future work will test generalizability across fields like biomedical research, physics, or social sciences.
- **Model Comparisons:** Evaluating other models such as Claude, Gemini, Mixtral, or LLaMA would provide deeper insight into the relationship between model architecture and review quality.
- **Long-Context Handling:** Future pipelines may integrate document summarization or retrieval-augmented generation to handle longer documents without sacrificing structure or coherence.

8.3 Challenges

Our project also posed several key challenges:

Text Extraction. Parsing academic PDFs proved difficult due to inconsistent formatting, multi-column layouts, figures, and mathematical notations. Extracting clean and structured input for LLMs remains a significant bottleneck.

Prompt Design. LLMs are not inherently trained to write structured academic reviews. Generic prompts often yielded shallow or loosely structured content. Iterative prompt engineering was necessary to achieve the desired tone and organization.

Evaluation Limitations. There is no universally accepted gold standard for evaluating review quality. Traditional metrics like ROUGE and BLEU capture surface-level similarity but miss deeper analytical value. This motivated our use of semantic metrics (BERTScore, SBERT) and a human evaluation component.

Baseline Limitations. Baseline reviews generated from generic prompts frequently omitted entire sections, such as weaknesses or suggestions, making it difficult to perform one-to-one content comparisons.

8.4 Final Remarks

Our system demonstrates that structured prompting can turn LLMs into capable assistants for academic review generation. While not a replacement for human experts, such systems can support peer review workflows, help early-career researchers receive feedback, and increase access to critical scientific commentary. We view this as a step toward more structured, rigorous, and human-aligned applications of generative AI in scholarly domains.

References

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

Pengfei Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A. Smith. 2022. From prompt engineering to prompt programming: Tools and practices. https://arxiv.org/abs/2211.01910. ArXiv preprint arXiv:2211.01910.

OpenAI. 2023. GPT-4 Technical Report. OpenAI Technical Report.

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*, pages 311–318. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.