PaperFormer: A Citation-Graph Enhanced Language Model for Scientific Applications

Anonymous ACL submission

Abstract

The rapid expansion of scientific literature presents significant challenges in navigating, understanding, and utilizing scholarly knowledge effectively. To address this, we introduce PaperFormer, a citation-network-aware Language Model designed to enhance scientific tasks by incorporating citation graph information. PaperFormer augments a base model with additional specialized weights to effectively process and analyze research papers within their citation contexts. To support this research, we also release a novel dataset[1] comprising approximately 10K papers, 42K reviews and rebuttals and 200K citation relationships. Our model undergoes pre-training on the Semantic Scholar Network (SSN) dataset and is evaluated across three tasks: causal language modeling, paper summarization, and automated review generation. Experimental results demonstrate that PaperFormer outperforms the state-of-the-art model in the paper summarization task and surpasses the base model in review generation. To foster further research, we open-source our models and the review-citations dataset, enabling broader adoption and extension of our work.

1 Introduction

The exponential rise in scientific publications has created a pressing need for more efficient methods to comprehend and analyze scholarly knowledge. Unlike general textual domains, scientific literature is highly structured, rich in domain-specific terminology, and deeply interconnected through citation networks that shape the context and significance of research contributions. Traditional approaches—such as keyword-based retrieval and abstract-based summarization—often fall short in capturing these relationships, limiting their ability to provide a comprehensive understanding of a paper's impact and relevance.

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable success in text generation tasks, yet existing models primarily focus on isolated documents without leveraging the broader citation landscape. Prior efforts in citation-aware modeling have largely been restricted to using abstracts of referenced papers, which may not provide insightful context for understanding a target paper. This limitation underscores the need for full-text processing and citation-aware modeling, enabling models to contextualize research within the evolving scholarly discourse. Two key tasks—summarization and review generation—serve as foundational steps toward this goal. Summarization distills the core contributions of a paper while integrating insights from its citation network, whereas review generation assists in the peer-review process by identifying strengths, limitations, and relevance within a broader research context. These tasks not only enhance literature comprehension but also alleviate the growing burden on human reviewers, enabling more efficient and insightful scholarly assessments.

To bridge this gap, we introduce PaperFormer, a citation-network-aware LLM that enhances a base model with additional specialized weights, enabling it to process and analyze scientific papers within their citation contexts. We use Llama 3.2-1B as the base model, leveraging its strong language modeling capabilities and extended context length of 128K tokens. Our approach incorporates citation-aware augmentations, allowing the model to integrate contextual information from reference papers and generate more informed and nuanced summaries, reviews, and rebuttals.

To support this research, we release a novel dataset comprising review and rebuttal comments on submitted papers and their citation relationships. Unlike prior datasets that primarily focus on summarization—where models extract key information from a single document—our dataset enables a new

[1] Will be released after double-blind peer reviews.

task: generating paper reviews and rebuttals. This task requires a deeper understanding of a paper's claims, its limitations, and its relationship to prior research, necessitating a model that can analyze both the target paper and its citation network. By integrating full-text processing with citation-aware modeling, our dataset and approach pave the way for more advanced AI-assisted scholarly workflows, making citation-aware LLMs a significant step forward in scientific text processing.

Our contributions are as follows:

- We design PaperFormer that augments the Llama 3.2-1B model with citation-network aware weights, enabling the model to integrate contextual information from full paper texts and their cited documents.
- We compile a comprehensive dataset of approximately 10K main papers, 42K review and rebuttal comments, and 200K citation relationships to facilitate research in citation-aware applications.
- We conduct extensive experiments across three tasks—causal language modeling, paper summarization, and automated review generation which used an LLM-based scoring system, demonstrating significant performance improvements over the base model.

2 Related Work

2.1 Scientific Paper Summarization

In recent years, leveraging the citation graph structure of scientific literature has emerged as a promising approach to enhance paper summarization models. Traditional summarization techniques primarily focus on the content of individual documents, often overlooking the rich contextual information embedded within citation networks. Incorporating citation relationships provides a broader understanding of a paper's impact and relevance within its research community.

Qazvinian and Radev, 2008 pioneered the use of citation summaries for scientific article summarization. Their method aggregated citation sentences referencing a target paper to construct a summary from the perspectives of citing documents. However, their approach exhibited a limited scope, relying solely on citation contexts without integrating content from the source paper itself. This omission often resulted in fragmented or biased summaries, as the citations alone may highlight specific contributions without capturing the paper's complete narrative.

Building on this foundation, An et al., 2021 introduced the Citation Graph-based Summarization model (CGSum), which integrates information from both the source paper and its citation graph. The authors constructed the Semantic Scholar Network (SSN) dataset, consisting of 141,000 papers and 661,000 citation relationships, to train and evaluate their model. While their results demonstrated that incorporating citation networks significantly enhances the quality of generated summaries, the approach had notable limitations. Specifically, they relied solely on abstracts from neighboring papers, omitting the full text, which may have limited the contextual richness of the summaries. Additionally, their use of LSTM-based architectures for the encoder and decoder, though effective at the time, has since been surpassed by transformer-based models in both performance and efficiency.

More recently, Luo et al., 2023 introduced CitationSum, a citation-aware summarization framework that employs graph contrastive learning (GCL) to identify and integrate salient content from references, effectively capturing the varying relevance between source papers and their citations. The authors introduced the PubMedCite dataset, comprising approximately 192,000 biomedical papers and 917,000 citation relationships, to train and evaluate their model. Experimental results on the SSN and PubMedCite datasets demonstrated that CitationSum achieves state-of-the-art performance in scientific paper summarization.

Despite its strong performance, CitationSum has several limitations. First, the architecture is computationally inefficient, as it requires generating neighbor embeddings for each pair of target and neighbor papers separately. Second, the model compresses each neighbor's representation into a single vector, which may cause information loss by discarding important contextual details. Finally, similar to CGSum, CitationSum optimizes its output by maximizing ROUGE (Lin, 2004) scores against reference papers, which can lead to overfitting to the metric rather than improving the actual summary quality.

Our proposed PaperFormer architecture directly addresses these limitations. Unlike CitationSum, PaperFormer shares neighbor representations across all citing papers, significantly improving efficiency. Additionally, it represents each neighbor using 512 distinct vectors, preserving richer

contextual information and mitigating information loss. Lastly, PaperFormer does not rely on ROUGE-based optimization, ensuring that its summaries are guided by semantic quality rather than arbitrary metric improvements.

2.2 Scientific Paper Review Generation

Several notable efforts have been made to automate and enhance the peer review process. For instance, Uban and Caragea, 2021 explored automatic review summary generation by leveraging neural language models, introducing a dataset of scientific papers and their reviews from NeurIPS conferences. Similarly, Wang et al., 2020 developed ReviewRobot, a system that assigns review scores and generates comments across multiple categories by constructing knowledge graphs from target papers and their references. More recently, the MARG framework proposed by D'Arcy et al., 2024 employs multiple GPT-4 instances to generate peer-review feedback through internal discussions, distributing paper text across agents to overcome input length limitations. Our work differs by integrating a citation-network-aware language model that considers not only the target paper but also its references. This enables our model to generate reviews that are contextually enriched by the relationships between papers and language understanding of a base large language model, offering insights that previous models, which often lack this extensive citation context, may not capture.

Creating datasets for peer review analysis has been pivotal for advancing research in this domain. Kang et al., 2018 introduced PeerRead, a dataset comprising 14.7K paper drafts, corresponding accept/reject decisions, and 10.7K textual peer reviews from venues like ACL, NIPS, and ICLR. D'Arcy et al., 2023 presented ARIES, a corpus containing review comments and their corresponding paper edits, facilitating research on revising scientific papers based on peer feedback. Additionally, the ORSUM dataset, as discussed by Zeng et al., 2024, encompasses 15,062 paper meta-reviews and 57,536 paper reviews from 47 conferences, aiming to support scientific opinion summarization. According to our knowledge, at the time of writing, there is no work that has utilized OpenReview reviews and paired them with their references. Our contribution extends this line of work by curating a dataset that not only includes papers and their reviews but also maps out the citation relationships among them. This enriched dataset supports the development and evaluation of models that leverage citation networks, enabling tasks such as generating reviews informed by a paper's citation context.

3 Methods

3.1 Generating Document Representations

Processing the full text of each paper within a dense citation network presents significant computational challenges, as a single paper may be referenced by multiple others. To avoid redundant computations and efficiently reuse paper representations, we generate concise yet informative embeddings using Llama 3.2-1B, which, with its extended context length, allows us to process an entire paper in a single pass. Specifically, we extract output vectors from the penultimate hidden layer, capturing a rich, contextualized representation that can be reused across different citation contexts without requiring repeated model inference.

Since each token in the document has a corresponding vector, directly using all of them would result in prohibitively large representations. To mitigate this, we retain only the last 512 output vectors, which, having attended to all preceding tokens, encapsulate a comprehensive contextual summary of the paper. Unlike Luo et al., 2023, CitationSum, which compresses references into a single pooled vector—typically derived from the abstract or selected salient content—our approach preserves finer-grained context from the full text, minimizing information loss and capturing a more detailed and nuanced representation of the paper's content.

3.2 Incorporating Citation Context in Model Architecture

Building upon the Llama 3.2-1B framework, our model incorporates mechanisms to integrate representations of neighboring (i.e., cited) documents, thereby enriching the context for the target document. The architecture comprises the following components:

- **Intra-Reference Refinement Block:** Each reference paper's 512-token representation is processed through a series of three identical attention blocks. Each block comprises an RMS normalization layer, a self-attention mechanism, another RMS normalization layer, and a fully connected feedforward network. This architecture mirrors the Gated Feedforward Network utilized in Llama 3.2-1B (Touvron

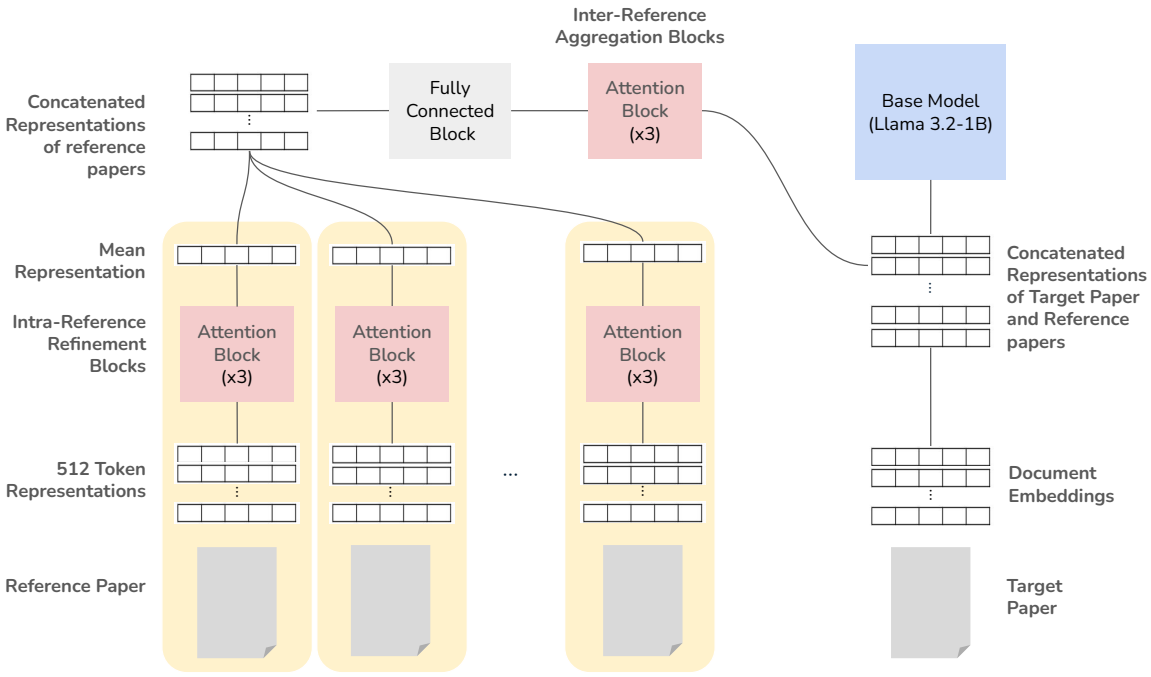

Figure 1: Overview of PaperFormer (with LLaMA 3.2-1B as Base Model) Architecture Flow: Each reference paper is encoded into a 512-token representation, refined by Intra-Reference Refinement Blocks, which are 3 repeated Attention Blocks, and then reduced to a single vector via mean pooling. These per-reference vectors are concatenated and further processed by an Inter-Reference Aggregation Block, which are again 3 repeated Attention Blocks, to capture cross-document relationships. Finally, the refined citation representations are concatenated with the target paper's embeddings and fed into the LLaMA 3.2-1B model for downstream language modeling, summarization, or review generation.

et al., 2023), where the feedforward component employs a gating mechanism to control the flow of information, enhancing model performance, whereas, RMS Norm, used in place of Layer Norm, improves training dynamics. Skip connections are incorporated to maintain gradient stability. The purpose of these blocks is to refine the token representations, enhancing their utility for subsequent language modeling tasks.

- **Aggregation of Neighbor Representations:** Post processing through the intra-reference refinement blocks, we compute the mean of the 512 token vectors for each reference paper, resulting in a single vector representation per document. These vectors are then concatenated to form a two-dimensional tensor representing all reference papers.

- **Inter-Reference Aggregation Block:** This block processes the concatenated neighbor representations, allowing the model to capture inter-document relationships and share information across the reference papers. The structure of this block parallels that of the intra-reference refinement block, facilitating effective information exchange among the neighbor representations.

- **Integration with Main Text:** The refined neighbor representations are concatenated with the embeddings of the main text of the target paper. This combined input is then fed into the Llama 3.2-1B model for language modeling tasks. By incorporating both the target document and its contextual citation information, the model is better equipped to understand and generate content that reflects the broader scholarly discourse.

This architecture enables the model to effectively integrate citation context, enhancing its performance in tasks such as language modeling, summarization, and review generation.

3.3 Training and Fine-Tuning

The training process involves an initial pre-training phase followed by fine-tuning:

- **Pre-Training:** The model was pre-trained on

the Semantic Scholar Network (SSN) dataset, which contains a large collection of scientific papers and their citation relationships. During pre-training, the base model, LLaMA 3.2-1B, remained frozen, with only the additional weights being updated. To address computational constraints, we limited the number of contextual neighbors to a maximum of eight. Additionally, due to resource limitations, we could not pre-train on the full text of every paper. Instead, for each paper, we selected a single 2048-token chunk from a random section and trained the model on that portion.

Through this process, the additional weights are intended to capture patterns in citation relationships, enabling the model to aggregate relevant contextual information and pass it to the base model. This approach aims to preserve the base model's general language understanding while allowing the added components to specialize in integrating citation context. It took an average of 40 hours with 1 Nvidia A100 40GB for pre-training.

- **Fine-Tuning:** The model is fine-tuned for scientific paper summarization or review generation, using either the SSN dataset or our curated training dataset, depending on the task. We use LoRA for finetuning, which adds low-rank adaptation layers to the model rather than modifying its original weights, making the process more efficient. During finetuning, the loss calculation is restricted to the summary or review tokens, rather than the entire prompt, to improve the model's performance on these specific tasks.

4 OpenReview Review-Rebuttal Dataset

Our dataset is curated to support the development and evaluation of citation-aware language models. It comprises approximately 10K papers, 42K reviews and rebuttals, and 200K citation relationships The dataset was formed by performing the following steps:

1. **Gathering Reviews and Rebuttals:** Using the OpenReview[2] API, we retrieve venues hosted on the platform, identifying 260 venues that contain papers with posted reviews. For each venue, we query the API to obtain detailed information on individual papers. The API returns this data in JSON format, including reviews and rebuttals, which we extract. In total, we collect 34,173 target papers at this stage.

2. **Finding the Citation Relations:** We used the Semantic Scholar API[3], which provides a comprehensive mapping of citation relationships between research papers. For each target paper, the API returns a list of referenced papers, from which we extract their arXiv IDs. Through this process, we successfully obtained reference information for 22,824 papers out of the 34,173 target papers.

3. **Collecting Full-Text Papers:** For the papers collected from Semantic Scholar, we filtered out those available on arXiv[4] and extracted their full text directly from their LaTeX source files. Since many scientific papers use LaTeX for formatting, extracting meaningful text requires handling various formatting commands, equations, and references. To achieve this, we used the open-source tool PyDetex[5], which processes LaTeX files and removes markup while preserving the text structure. This approach ensures that extracted content retains readability without unnecessary formatting artifacts. By using the original LaTeX sources, we minimize errors introduced by PDF-to-text conversion and obtain cleaner textual data for training and evaluation. However, since LaTeX source files need to be downloaded individually, some files are partially extracted or completely corrupted due to download errors or formatting inconsistencies. To filter out such cases, we exclude papers with extracted text containing fewer than 8,000 characters, as these are likely incomplete or incorrectly processed. We are then left with 10,432 main papers and 200k citation relations that have full texts.
Since publicly releasing this part of the dataset would require obtaining explicit permissions from the authors of each paper, we instead provide scripts[6] that allow researchers to extract the full texts themselves. These scripts automate the LaTeX processing and text extrac-

[2] OpenReview

[3] Once we obtain a list of target papers, we retrieve their referenced papers through the Semantic Scholar API

[4] arXiv

[5] PyDetex

[6] Will be released after double-blind peer reviews.

tion pipeline, ensuring reproducibility while respecting copyright and data ownership concerns.

Each entry in our dataset contains a paper's associated review(s) and rebuttal(s), and a list of references (with their corresponding full texts). We apply a 90/10 train/test split, resulting in 9,388 papers for training and 1,044 for testing. To enhance the review task dataset, we augment the training set by splitting multiple reviews per paper into individual training examples. This augmentation yields approximately 36K pairs of paper full texts and reviews.

5 Experiments

5.1 Evaluation Setup

We evaluate PaperFormer on three tasks: causal language modeling, paper summarization, and automated review generation. Performance is measured using perplexity for language modeling, ROUGE F1 (rouge-score==0.1.2 Pip package) for summarization, and alignment with ground-truth reviews for review generation.

For benchmarking, we compare PaperFormer against the standalone Llama 3.2-1B model fine-tuned using Low-Rank Adaptation (LoRA). We set the LoRA rank to 8, the scaling factor to 16, and apply a dropout rate of 0.1, following best practices. This evaluation framework allows us to assess the impact of incorporating citation context on model performance across different tasks.

5.2 LLM-as-a-Judge Evaluation

To compare the quality of reviews generated by PaperFormer and the baseline model, we employ an LLM-as-a-Judge approach. In this setup, an LLM is presented with two anonymized reviews—one from the baseline model and one from PaperFormer—along with the expert reviews of the paper. The LLM is prompted to select which review aligns more closely with the expert reviews, choosing between Review A, Review B, or Tie. To mitigate positional bias, the position of PaperFormer's review is randomized between A and B.

We use GPT-4o as the judge for this evaluation, following the prompting methodology outlined in Zheng et al., 2023. This approach has been shown to achieve approximately 80% agreement with human evaluations, making it a reliable proxy for comparative assessment. The LLM-as-a-Judge framework leverages the high consistency and scalability of LLM-based evaluations, reducing the cost and time required for human assessments while maintaining a strong correlation with expert judgments. Prior work demonstrates that LLMs can effectively assess text quality across various domains, particularly when provided with clear evaluation criteria and expert-validated reference texts. By structuring our evaluation to align with this framework, we ensure consistency, reproducibility, and scalability in our comparative analysis. The exact prompts used for evaluation are provided in the Appendix Figure 5 for transparency.

Model	Perplexity
Llama-3.2-1B	14.12
PaperFormer+Llama-3.2-1B	7.23
PaperFormer+Llama-3.2-1B*Finetuned*	**6.78**

Table 1: Causal Language Modelling Task results. All scores are for single runs.

5.3 Result Analysis and Discussion

1. **Causal Language Modeling (CLM):**
   In the causal language modeling task, we assess the models' performance using perplexity scores, where lower values indicate better predictive capabilities. The results are summarized in Table 1. As shown in Table 1, PaperFormer achieves a perplexity of 7.23, significantly outperforming the standalone Llama 3.2-1B model, which has a perplexity of 14.12. Further fine-tuning of PaperFormer leads to a slight improvement, reducing the perplexity to 6.78. These results demonstrate the effectiveness of incorporating citation context through the PaperFormer's additional weights in enhancing language modeling performance.

2. **Paper Summarization:**
   We evaluate PaperFormer against both extractive and abstractive summarization baselines, including SOTA models CGSum (An et al., 2021) and CitationSum (Luo et al., 2023). Extractive methods include LEAD (See et al., 2017), which selects the first L sentences; TextRank (Mihalcea and Tarau, 2004), a graph-based ranking algorithm; TransformerEXT ((Liu and Lapata, 2019)), a transformer encoder-based extractor; and BERTSUMEXT (Liu and Lapata, 2019), a BERT-based extractive model. Abstractive

methods include PTGEN+COV (See et al., 2017), which uses a pointer-generator network with coverage mechanisms; TransformerABS (Liu and Lapata, 2019), a transformer-based abstractive summarizer; and BERTSUMABS (Liu and Lapata, 2019), which builds on BERT for abstractive summarization. CGSum incorporates citation graphs for improved summarization, while CitationSum leverages BERT-based models with citation-aware mechanisms. For fair comparison, we report baseline scores directly from the CGSum and CitationSum papers, ensuring consistency in evaluation metrics and experimental setup. These baselines provide a strong benchmark for assessing PaperFormer's effectiveness in generating citation-aware, context-rich scientific summaries.

Table 2 presents ROUGE F1 scores for the paper summarization task. PaperFormer achieves state-of-the-art performance, surpassing CitationSum+BERT+PubMedBERT and all other baselines in ROUGE-1 (47.85) and ROUGE-2 (19.81), while also improving on ROUGE-L. Notably, PaperFormer outperforms Llama-3.2-1B-Instruct Finetuned, demonstrating the benefit of incorporating citation context in summarization.

The improvements suggest that leveraging full-text representations and reference embeddings allows PaperFormer to capture more relevant information than models relying solely on the abstract or selected content. This finding aligns with our hypothesis that citation-aware document representations contribute to more informative and contextually rich summaries.

3. **Review Generation:**
The review generation task results (Table 3) indicate that PaperFormer produces more informative and aligned reviews compared to the baseline models. It achieves the highest scores across all ROUGE metrics (ROUGE-1: 41.11, ROUGE-2: 14.06, ROUGE-L: 23.80), outperforming Llama-3.2-1B-Instruct and its fine-tuned variant.

Additionally, the LLM-as-a-Judge evaluation (Figure 1) supports these findings. PaperFormer's generated reviews were preferred 538 times over the baseline's 480 times, with

Metrics	**R-1**	**R-2**	**R-L**
LEAD	28.29	5.99	24.84
TextRank	36.36	9.67	32.72
TransformerEXT	43.14	13.68	38.65
BERTSUMEXT	42.41	13.10	37.97
BERTSUMEXT+	44.28	14.67	39.77
PTGEN+COV	42.84	13.28	37.59
Concat Nbr.Summ	43.05	13.53	37.97
TransformerABS	37.78	9.59	34.21
+*Copy*	41.22	13.31	37.22
BERTSUMABS	43.73	15.05	39.46
BERTSUMABS+	43.73	15.05	39.46
Concat Nbr.Summ	43.45	14.89	39.27
CGSUM	44.28	14.75	39.76
CitationSum+BERT	44.72	15.03	40.12
+ PubMedBERT	45.01	15.18	**40.59**
Llama-3.2-1B-Instruct	29.49	10.24	17.65
Llama-3.2-1B-Instruct Finetuned	47.21	19.35	27.60
PaperFormer+Llama-3.2-1B-Instruct	**47.85**	**19.81**	28.14

Table 2: ROUGE F1 Results on the SSN dataset for Inductive settings on the Paper Summarization Task, for single run. R-1, R-2, and R-L refer to ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. Best scores are marked as **Bold**.

Type	Reviews		
Metrics	**R-1**	**R-2**	**R-L**
Llama-3.2-1B-Instruct	28.84	6.07	14.46
Llama-3.2-1B-Instruct Finetuned	40.02	13.62	23.33
PaperFormer+Llama-3.2-1B-Instruct	**41.11**	**14.06**	**23.80**

Table 3: ROUGE F1 Results on our dataset for Reviews and Rebuttals Generation Task, for single run. R-1, R-2, and R-L refer to ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. Best scores are marked as **Bold**.

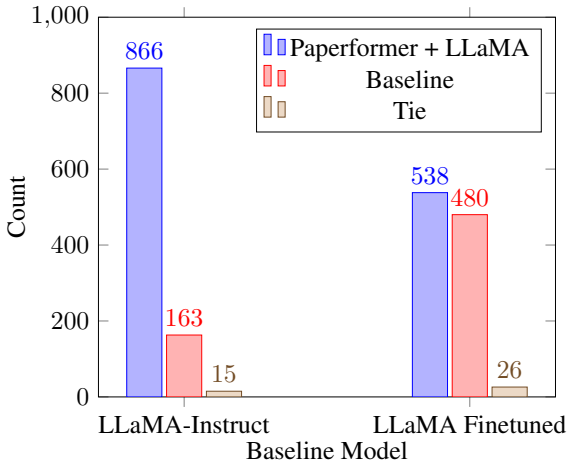

Figure 2: Performance Comparison of Paperformer + LLaMA with baseline models as LLaMa-Instruct and LLaMa Finetuned

26 cases marked as ties. This suggests that reviews incorporating citation-aware representations align more closely with expert assessments. The improvements in both ROUGE and LLM-based evaluations highlight the importance of citation integration for generating better scientific reviews and rebuttals.

6 Conclusion and Future Work

In this study, we introduced PaperFormer, a citation-network-aware Large Language Model designed to enhance the processing and analysis of scientific papers within their citation contexts. By integrating the Llama 3.2-1B model with additional specialized weights, PaperFormer effectively incorporates citation information, enabling a more comprehensive understanding of scientific literature.

Our evaluations across three tasks—causal language modeling, paper summarization, and automated review generation—demonstrate the effectiveness of PaperFormer in leveraging citation context to improve performance. Notably, PaperFormer outperforms the standalone Llama model in these tasks, highlighting the benefits of incorporating citation information.

Future work will explore scaling the model using larger base LLMs to improve document understanding and assess how performance scales with model size. Beyond summarization and review generation, we aim to extend the model to broader scientific NLP tasks such as citation intent classification and claim verification. Architecturally, we plan to move beyond the current 512-token representation by enabling full-text processing, leveraging memory-efficient transformers or sparse attention mechanisms. Instead of limiting input to 2048 tokens, incorporating entire documents and references could improve context comprehension and the quality of generated outputs.

By advancing the development of citation-aware language models, we aim to contribute to the improvement of tools and methodologies that support the scientific community in managing and understanding scientific literature more effectively.

7 Limitations

While PaperFormer demonstrates significant advancements in citation-aware language modeling for scientific applications, several limitations remain. To maintain computational efficiency, we limit the number of citation neighbors to a maximum of eight. While this allows for feasible processing of citation graphs, it inevitably excludes additional contextual references that could be informative, especially for highly cited papers. Moreover, our evaluation relies predominantly on quantitative metrics such as ROUGE scores and LLM-as-a-Judge assessments. Although these provide useful insights into model performance, they may not capture all aspects of quality in generated summaries or reviews—particularly the subtleties of scholarly critique and discourse. More comprehensive human evaluations may be required to fully assess the practical impact and reliability of the generated outputs.

8 Ethics Statement

Our model introduces the review generation task primarily as a way to evaluate how well language

models integrate citation information, rather than as a tool to replace human reviewers. While LLMs can surface relevant insights and highlight key areas of interest, they are not reliable for conducting peer reviews due to well-documented limitations such as hallucinations, lack of deep domain understanding, and inconsistencies in reasoning. Given these challenges, expert oversight remains essential for scientific critique. Instead of serving as an authoritative evaluator, our model should be viewed as a benchmark for assessing citation-aware text generation. To ensure responsible use, we advocate for transparency in model outputs and recommend that any automated assessments be accompanied by clear disclaimers outlining their limitations. Ultimately, human judgment is critical in maintaining the rigor and integrity of the peer review process, with AI functioning as a means to study citation-aware language modeling rather than a direct substitute for reviewers.

Acknowledgments

ChatGPT was used to enhance the coherence and cohesiveness of the text in this paper.

References

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2021. Enhancing scientific papers summarization with citation graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12498–12506.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *Preprint*, arXiv:1908.08345.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Citationsum: Citation-aware graph contrastive learning for scientific paper summarization. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1843–1852, New York, NY, USA. Association for Computing Machinery.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, page 689–696, USA. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ana Sabina Uban and Cornelia Caragea. 2021. Generating summaries for scientific paper review. *Preprint*, arXiv:2109.14059.

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*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. *Preprint*, arXiv:2305.14647.

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 *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Appendix

A.1 Summary Generation

Figure 4(a) shows the abstract of a paper from the dataset, and Figure 4(a) shows the Summary generated by the PaperFormer for the same paper.

A.2 Review Generation

Figure 5 shows a sample of a review comment generated by the PaperFormer.

A.3 Prompts used for LLM-as-a-Judge evaluation

Figure 6(a) and 6(b) show the System and User Prompt, respectively, used for GPT-4o to evaluate the baseline and PaperFormer reviews.

quantum coherence plays a central role in various research areas . the $ l_1$-norm of coherence is one of the most important coherence measures that are easily computable , but it is not easy to find a simple interpretation . we show that the $ l_1$-norm of coherence is uniquely characterized by a few simple axioms , which demonstrates in a precise sense that it is the analog of negativity in entanglement theory and sum negativity in the resource theory of magic - state quantum computation . we also provide an operational interpretation of the $ l_1$-norm of coherence as the maximum entanglement , measured by the negativity , produced by incoherent operations acting on the system and an incoherent ancilla . to achieve this goal , we clarify the relation between the $ l_1$-norm of coherence and negativity for all bipartite states , which leads to an interesting generalization of maximally correlated states . surprisingly , all entangled states thus obtained are distillable . moreover , their entanglement cost and distillable entanglement can be computed explicitly for a qubit - qudit system .

(a) Abstract of a paper

1) we propose an axiomatic characterization of the $ l_1 $-norm of coherence, which reveals a precise analogy to the negativity in entanglement theory and the sum negativity in the resource theory of magic - state quantum computation. 2) we provide an operational interpretation of the $ l_1 $-norm of coherence as the maximum entanglement, measured by the negativity, created by incoherent operations acting on the system and an incoherent ancilla. 3) we clarify the relation between the $ l_1 $-norm of coherence and negativity for bipartite states. we also propose pairing states as a generalization of maximally correlated states and as a bridge for connecting coherence theory and entanglement theory. surprisingly, the asymptotic exact ppt entanglement cost of any pairing state is equal to the logarithmic negativity. moreover, such states are distillable whenever they are entangled, and their entanglement cost and distillable entanglement can be computed explicitly for a qubit - qudit system.

(b) Summary of the above paper generated by PaperFormer

Figure 3: Summary Generation

Review

Summary
The paper proposes a method for simultaneously CSR and spherical mapping, which is efficient in seconds. The approach is based on a recurrent strategy to learn large deformations from the sphere to the white and pial surfaces while maintaining architectural simplicity. The method is evaluated on a dataset of infants from 0 to 12 months, demonstrating its efficacy in improving mesh regularity and reducing geometric errors.

Soundness
3

Presentation
3

Contribution
3

Strengths
1. The proposed method is efficient in seconds, which is a significant improvement over traditional methods that require hours or even days for CSR.
2. The method is evaluated on a dataset of infants from 0 to 12 months, demonstrating its efficacy in improving mesh regularity and reducing geometric errors.
3. The paper is well-written and easy to follow.

Weaknesses
1. The method is based on a recurrent strategy to learn large deformations from the sphere to the white and pial surfaces while maintaining architectural simplicity. However, the recurrent strategy may be computationally expensive and may not be suitable for large-scale datasets.
2. The method is evaluated on a dataset of infants from 0 to 12 months, which may not be representative of the general population. It would be interesting to see the results on a larger dataset, such as the MNI-CAT dataset, which has a larger number of subjects and a more diverse range of ages.

Questions
1. The method is based on a recurrent strategy to learn large deformations from the sphere to the white and pial surfaces while maintaining architectural simplicity. However, the recurrent strategy may be computationally expensive and may not be suitable for large-scale datasets. How does the computational cost of the method compare to other methods?
2. The method is evaluated on a dataset of infants from 0 to 12 months, which may not be representative of the general population. It would be interesting to see the results on a larger dataset, such as the MNI-CAT dataset, which has a larger number of subjects and a more diverse range of ages.

Flag_for_ethics_review
- No ethics review needed.

Rating
6

Confidence
3

Code_of_conduct
Yes

Figure 4: Sample review comment generated by PaperFormer

You are the chair of a conference who is reviewing the quality of reviews. Please act as an impartial judge and evaluate the quality of the reviews provided by two reviewers comparing them with the expert reviews displayed below. You will be given a series of expert reviews, followed by reviewer A's answer, and reviewer B's answer. Your job is to evaluate which reviewer's answer is better. Begin your evaluation by comparing both reviewers' answers with the expert reviews. See how well each section of the reviews matches with the sections in the expert reviews. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: \"[[A]]\" if assistant A is better, \"[[B]]\" if assistant B is better, and \"[[C]]\" for a tie."

(a) System Prompt

Here are the list of expert reviews:
{expert_reviews}

Now choose which of the two reviews aligns better with the expert reviews:
[The Start of Reviewer A's Answer] Review A:
{review_a}
[The End of Reviewer A's Answer]
[The Start of Reviewer B's Answer] Review B:
{review_b}
[The End of Reviewer B's Answer]

(b) User Prompt

Figure 5: Prompts used for LLM-as-a-Judge evaluation