Can LLMs Identify Critical Limitations within Scientific Research?
A Systematic Evaluation on AI Research Papers

Zhijian Xu[*][Y]   **Yilun Zhao**[*][Y]   **Manasi Patwardhan**[T]
Lovekesh Vig[T]   **Arman Cohan**[Y]

[Y] Yale University   [T] TCS Research

Abstract

Peer review is fundamental to scientific research, but the growing volume of publications has intensified the challenges of this expertise-intensive process. While LLMs show promise in various scientific tasks, their potential to assist with peer review, particularly in identifying paper limitations, remains understudied. We first present a comprehensive taxonomy of limitation types in scientific research, with a focus on AI. Guided by this taxonomy, for studying limitations, we present LIMITGEN, the first comprehensive benchmark for evaluating LLMs' capability to support early-stage feedback and complement human peer review. Our benchmark consists of two subsets: LIMITGEN-Syn, a synthetic dataset carefully created through controlled perturbations of high-quality papers, and LIMITGEN-Human, a collection of real human-written limitations. To improve the ability of LLM systems to identify limitations, we augment them with literature retrieval, which is essential for grounding identifying limitations in prior scientific findings. Our approach enhances the capabilities of LLM systems to generate limitations in research papers, enabling them to provide more concrete and constructive feedback.

🤗 **Data** `yale-nlp/LimitGen`
○ **Code** `yale-nlp/LimitGen`

Figure 1: Overview of the research: the limitation generation task and three research questions investigated.

1 Introduction

Peer review plays a crucial role in ensuring the quality and integrity of scientific research. However, it is often a time-consuming and expertise-intensive process, posing significant challenges, especially as the volume of published papers continues to grow. Recent advancements in large language models (LLMs) have demonstrated remarkable capabilities across a variety of scientific tasks, such as answering questions about scientific papers (Xu
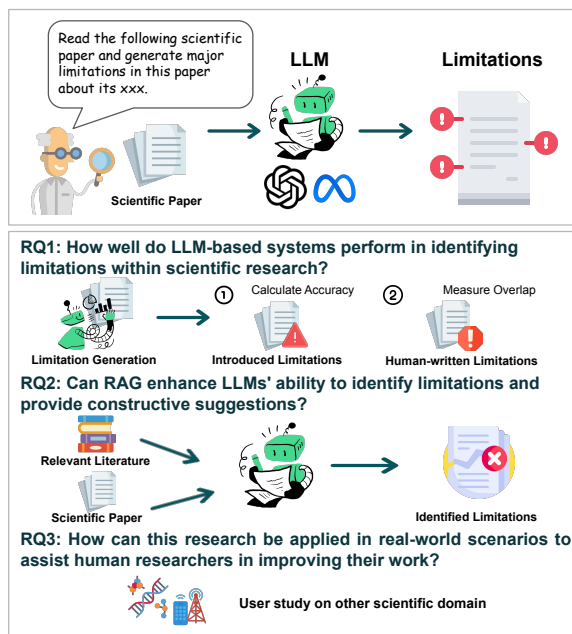
et al., 2024), writing scientific papers (Chamoun et al., 2024; Lu et al., 2024), retrieving related works (Ajith et al., 2024; Press et al., 2024), improving idea generation (Wang et al., 2024; Zhou et al., 2024b; Si et al., 2024), and generating code to perform data-driven discovery (Huang et al., 2024; Tian et al., 2024). Meanwhile, there is increasing interest in exploring the potential of LLMs to assist with or generate peer reviews (Liang et al., 2024; Liu and Shah, 2023; D'Arcy et al., 2024; Lou et al., 2024), which can be used to provide quick and early feedback to researchers and potentially alleviate some of the burdens associated with traditional review processes.

High-quality reviews are supposed to pinpoint the limitations of a paper and provide concrete, actionable suggestions, assisting researchers in improving their work. Existing benchmarks for peer-review generation collect papers and their corre-

---

[*]Equal Contributions.

sponding entire reviews from AI conferences (Du et al., 2024; Yu et al., 2024; Tan et al., 2024), but these benchmarks generally do not emphasize the importance of limitation identifications. Instead, they compare the overall quality of LLM-generated versus human-written reviews (Liang et al., 2024; Yu et al., 2024) and assess adherence to conference guidelines (Tyser et al., 2024). As opposed to other aspects of review generation, such as summary, strengths, syntactic or structural errors or request for elaboration, identification of limitations is the key aspect to facilitate future technical enhancement of the work and it is of utmost importance from the research growth point of view.

To address this gap, we present the first in-depth study and evaluation of LLM systems in identifying the limitations of scientific papers. To do so, we first provide a comprehensive taxonomy of types of limitations in scientific fields of study, with focus on AI.[1] Then guided by this taxonomy, we propose LIMITGEN-Syn, a synthetic benchmark focusing on various categories of limitations. LIMITGEN-Syn systematically introduces controlled perturbations to high-quality papers to create scenarios where specific limitations are present. These perturbations include selective removal of crucial information such as experimental details, inadequate evaluation metrics, omission of key baseline comparisons, and constraints on datasets or methodologies. By carefully controlling these modifications, we can reliably evaluate how well LLM agents detect different types of limitations. In addition, LIMITGEN-Syn also allows capturing suggestions on how to resolve the identified limitation.

To assess whether our taxonomy and the synthetic benchmark can effectively capture the diverse categories of limitations identified by humans in real-world peer review settings, we then collect human-written limitations from ICLR 2025 submissions as LIMITGEN-Human. We chose ICLR 2025 to mitigate contamination and also because ICLR peer reviews are often of high-quality due to the public nature of reviews and an extensive rebuttal process. Together, these two datasets form a comprehensive benchmark for advancing identification of limitations in papers.

Limitation identification is a knowledge-intensive task, requiring years of expertise and staying current with rapidly evolving literature. In

terms of modeling, in such settings, retrieval plays a crucial role as it is challenging to keep LLM agents up-to-date in rapidly evolving fields. While Retrieval-augmented generation (RAG) has been applied to enhancing scientific workflow, such as conducting literature reviews (Agarwal et al., 2024; Asai et al., 2024) and answering domain-specific questions (Xu et al., 2024; Skarlinski et al., 2024), it has not yet been explored in the context of peer-reviews. RAG also simulates how human identify limitations of papers by implicitly or explicitly referring to existing body of (often recent) literature and thus facilitate in grounding the generated limitations in existing scientific findings. We enhance limitation generation by leveraging RAG techniques. Specifically, we prompt LLMs to query the Semantic Scholar API to retrieve papers related to the one under review, extracting relevant content to enrich their domain understanding. Our results demonstrate that incorporating RAG improves the ability of LLM systems to generate limitations in research papers, providing more concrete feedback.

We summarize our contributions as follows:

- We propose LIMITGEN, a comprehensive benchmark specifically designed to assess the ability of models to identify and address limitations in scientific research, with a reliable and systematic evaluation framework.

- We evaluate the performance of LLMs and agent-based systems in identifying limitations and demonstrate their shortcomings in providing constructive and actionable feedback.

- We explore the potential of RAG in review generation, demonstrating its ability to improve limitation identification and generate more contextually relevant and actionable suggestions.

2 Related Work

2.1 Peer-Review Generation

Recent advances in LLMs have significantly influenced scientific research, offering tools to streamline and enhance researchers' workflows across various stages of the scientific pipeline (Xu et al., 2024; Lu et al., 2024; Ajith et al., 2024; Wang et al., 2024; Zhou et al., 2024b; Si et al., 2024; Tian et al., 2024). Researchers have also extensively explored the potential of LLMs in automated peer review generation, employing various approaches such as guiding LLMs with single prompts (Liang et al.,

---

[1]We chose AI as this is the field we are familiar with. In §6.3 we also perform a user study to how our findings generalize to other domains.

2024), adopting two-stage review generation frameworks with question-guided prompts (Gao et al., 2024), and leveraging multi-agent systems (D'Arcy et al., 2024). Other studies simulate the complete review process as a multi-round dialogue (Tan et al., 2024). However, some research has found that LLM-generated reviews often suffer from generic and paper-unspecific content (Du et al., 2024), are seldom entirely accurate, lack critical analysis, and fail to provide technical details (Zhou et al., 2024a). During the peer review process, identifying limitations is a crucial task as it helps highlight weaknesses in a study, guiding authors toward improvements and fostering scientific progress. Current research primarily focuses on generating the entire review. Du et al. (2024) collected human and LLM-generated reviews, each annotated by experts with fine-grained deficiency labels and explanations. Tyser et al. (2024) compares generated reviews of papers with and without inserted errors by evaluating review scores. Other studies have constructed several exceptionally short computer science papers, each with an inserted error (Liu and Shah, 2023), or focused on identifying weaknesses within a single paragraph rather than an entire paper (Chamoun et al., 2024). Lou et al. (2024) extracted human-written weaknesses from peer reviews. However, these studies do not thoroughly evaluate whether LLM systems can effectively detect specific limitations in scientific research. In this work, we present a comprehensive benchmark to evaluate models' ability to identify and address limitations in AI research papers, comprising a synthetic subset created via controlled perturbations and a set of human-written limitations.

2.2 Retrieval Augmented Generation

Despite showing promise in various tasks, LLMs face significant challenges when adopted to specialized domains, including hallucinations (Mallen et al., 2023; Mishra et al., 2024), conflict between outdated pre-training data and latest domain knowledge (Kasai et al., 2024), and lack of transparent attribution (Ye et al., 2024). Retrieval augmented generation that integrates external knowledge has emerged as a pivotal strategy to address these limitations (Lewis et al., 2020; Shuster et al., 2021; Izacard et al., 2023), enabling LLMs to produce more accurate and context-aware outputs.

Recent studies use proprietary LLMs with external APIs (*e.g.,* Semantic Scholar API & Google Search API) (Agarwal et al., 2024; Skarlinski et al., 2024; Chamoun et al., 2024) or develop new methodologies to train specialized open models (Asai et al., 2024) for tasks such as scientific literature review. Furthermore, multiple-round retrieval-enhanced reasoning methods have been developed to improve retrieval effectiveness (He et al., 2022; Shao et al., 2023; Jiang et al., 2023; Chen et al., 2024). In this work, we introduce a novel approach that incorporates literature retrieval into the limitation generation process, enabling LLMs to utilize domain knowledge and produce more constructive feedback.

3 LIMITGEN Benchmark

This section discusses the task formulation of LIMITGEN and details the data construction process used to curate its two subsets.

3.1 Task Formulation

We formally define the task of limitation generation in the context of LLMs as follows: Given: (1) a scientific paper, which may either contain a major limitation explicitly introduced (*i.e.,* LIMITGEN-Syn discussed in §3.3), or exhibit limitations previously identified during peer review by human reviewers (*i.e.,* LIMITGEN-Human discussed in §3.4); and (2) an aspect of limitations, which serves as a focus point for the LLM to evaluate a specific dimension of the paper's quality. The LLM is tasked with generating the limitation for the given paper, reflecting its quality with respect to the specified aspect.

3.2 Desiderata and Taxonomy of Limitations

The identification and categorization of limitations in scientific research require careful consideration of what constitutes a meaningful limitation. Through our pilot analysis of peer reviews, we establish several key desiderata that guide our taxonomy of limitations. First, a research limitation should represent a substantive constraint or weakness that impacts the validity, generalizability, or reliability of the study's findings. These constraints may arise from methodological choices, resource limitations, or gaps in current scientific understanding. Importantly, limitations should be distinguished from superficial critiques of presentation style. Second, limitations should be actionable - they should point to specific aspects of the research that could be improved through concrete steps. This ensures that identifying limitations serves a constructive purpose in advancing scientific knowledge, rather than simply highlighting un-

Aspect	Limitation Subtype	Definition and Corresponding Data Example
Methodology	Low Data Quality	The data collection method is unreliable, potentially introducing bias and lacking adequate preprocessing (Figure 4)
	Inappropriate Method	Some methods in the paper are unsuitable for addressing this research question and may lead to errors or oversimplifications (Figure 5)
Experimental Design	Insufficient Baselines	Fail to evaluate the proposed approach against a broad range of well-established methods (Figure 6)
	Limited Datasets	Rely on limited datasets, which may hinder the generalizability and robustness of the proposed approach (Figure 7)
	Inappropriate Datasets	Use of inappropriate datasets, which may not accurately reflect the target task or real-world scenarios (Figure 8)
	Lack of Ablation Studies	Fail to perform an ablation study, leaving the contribution of a certain component to the model's performance unclear (Figure 9)
Result Analysis	Limited Analysis	Rely on insufficient evaluation metrics, which may provide an incomplete assessment of the model's overall performance (Figure 10)
	Insufficient Metrics	Offer insufficient insights into the model's behavior and failure cases (Figure 11)
Literature Review	Limited Scope	The review may focus on a very specific subset of literature or methods, leaving out important studies or novel perspectives (Figure 12)
	Irrelevant Citations	Include irrelevant references or outdated methods, which distract from the main points and undermine the strength of conclusions (Figure 13)
	Inaccurate Description	Provide an inaccurate description of existing methods, which can hinder readers' understanding of the context and relevance of the proposed approach (Figure 14)

Table 1: The types of scientific paper limitations included in the LIMITGEN-Syn subset.

avoidable constraints. For instance, a limitation regarding insufficient experimental validation should suggest specific additional experiments that would strengthen the work. Third, limitations should be grounded in established scientific principles and practices within the relevant domain. This requires domain expertise to properly identify and articulate limitations that reflect meaningful departures from best practices or gaps in scientific rigor. For instance, appropriate evaluation metrics for each task are well-known within each subfield. Based on these desiderata and our analysis of peer review comments from top AI conferences, we categorize research limitations into four primary aspects (Table 1): *(i) Methodological Limitations* focus on the fundamental approaches and techniques employed in the research. These include issues such as inappropriate choice of methods, unstated assumptions that may not hold, and problems with data quality or preprocessing that could introduce bias. Such limitations directly impact the validity of the research findings. *(ii) Experimental Design Limitations* encompass weaknesses in how the research validates its claims. This category includes insufficient baseline comparisons, limited datasets that may not represent the full problem space, and

lack of ablation studies to isolate the contribution of different components. These limitations affect the reliability and reproducibility of results. *(iii) Results and Analysis Limitations* relate to how findings are evaluated and interpreted. This includes using inadequate evaluation metrics that may not capture important aspects of performance, insufficient error analysis, and lack of statistical significance testing. These limitations impact the strength and generalizability of conclusions. *(iv) Literature Related Limitations* focus on how the research connects to and builds upon existing work. This includes missing citations of relevant prior work, mischaracterization of existing methods, and failure to properly contextualize contributions within the broader research landscape. These limitations affect both the novelty claims and the proper attribution of ideas. This taxonomy guides our creation of the LIMITGEN benchmark by ensuring we systematically evaluate different types of limitations that matter for scientific rigor. For each aspect, we identify specific subtypes of limitations that commonly appear in peer reviews and can be reliably assessed. The taxonomy also informs our evaluation criteria, as different types of limitations may require different forms of evidence and levels of

domain knowledge to properly identify.

3.3 LIMITGEN-Syn Subset Collection

Source Paper Collection We collect scientific papers from arXiv under the "Computation and Language" category, focusing on those released between March 1, 2024, and May 31, 2024, a period likely outside the pretraining data cut-off for most current LLMs. This selection helps minimize potential data memorization issues that affect model evaluation. To extract content, we use the tool[2] by Lo et al. (2020), which converts LaTeX source files into JSON format, capturing elements including the title, abstract, main sections, and appendix of each paper. In total, we compile an initial pool of 1,408 NLP papers for further annotation. We exclude papers that do not focus on experimental work, such as surveys, position papers, and dissertations, as these lack the experimental designs required for our analysis. Additionally, papers of insufficient quality are omitted to ensure that the introduced limitation represents the most critical issue in each paper. This filtering process led us to 500 papers.

Example Curation Following the taxonomy in Table 1, we design perturbation pipelines for each limitation subtype. For each paper, human experts determine the applicable perturbations and then apply all suitable perturbations accordingly. The annotators identify all the relevant sections in the paper based on the perturbation type. For each selected section, we employ GPT-4o to perturb the content according to the specific definitions and guidelines, such as removing relevant details or replacing a particular dataset. The prompts are provided in Figure 4 to Figure 14. Alongside each perturbation, we generate a brief description of the introduced limitation as the ground truth, which will serve as a reference for later evaluations.

Human Expert Validation To guarantee the reliability of our LIMITGEN-Syn dataset, each annotated example is evaluated by a human annotator based on the following criteria: (1) The text within the paper must be grammatically correct and maintain clarity. (2) The introduced limitation must genuinely impact the quality and represent the most critical issue in the given aspect. (3) The generated ground truth limitation should clearly articulate the problem and be reasonable. Validators are tasked

Property (`avg.`/`max`)	Value
LIMITGEN-Syn	
Scientific Paper Word Length	5,201.46 / 58,788
Limitation Word Length	34.45 / 81
Paper Number	500
Example Number	1,000
LIMITGEN-Human	
Scientific Paper Word Length	8,255.38 / 1,8910
Limitation Word Length	61.97 / 795
Number of Limitations per Paper	6.05 / 20
Paper Number	1,000

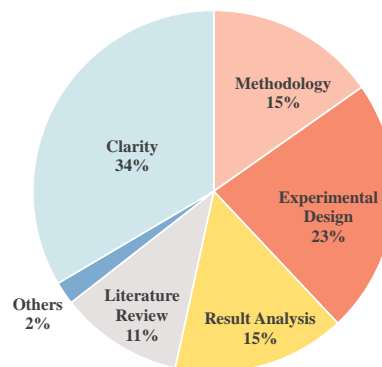Table 2: Data statistics of the LIMITGEN benchmark.

Figure 2: The aspect distribution of human-written limitations in LIMITGEN-Human.

with revising or removing examples that do not meet these standards. In practice, from 500 papers, a total of 1,000 examples were retained, including 112 that were revised by human annotators. We provide the details of annotators involved in dataset construction in Table 7.

3.4 LIMITGEN-Human Subset Collection

To assess whether our taxonomy and the synthetic benchmark can effectively capture the diverse categories of limitations identified by humans in real-world peer review settings, we then collect human-written limitations from ICLR 2025 submissions.

We specifically focus on the weaknesses sections of each paper's reviews and break them down into itemized limitations.

To ensure quality, we use GPT-4o to exclude weaknesses that are too short (fewer than 20 words) or lack substantive suggestions, and then categorize the remaining limitations. The prompt is provided in Figure 16.

We retained only the limitations related to methodology, experimental design, result analysis, and literature review, considering them as ground

truth. We collect a total of 9,844 papers and randomly sample 1,000 of them for experimentation.

3.5 Data Statistics

Table 2 illustrates the data statistics of our benchmark. Figure 2 presents the detailed aspect distributions of the LIMITGEN-Syn subset. The full LIMITGEN benchmark consists of 2,000 examples and encompasses a diverse range of aspect types commonly found in paper limitations.

4 LIMITGEN Evaluation Protocol

Evaluating the quality of limitations generated by LLMs is inherently challenging due to the subjective and nuanced nature of research critique. Such assessments typically require expert-level judgment, making human evaluation labor-intensive. Moreover, comparing generated limitations with ground-truth ones is non-trivial, as valid limitations may differ in phrasing or granularity. These challenges motivate the careful design of our evaluation protocol to ensure both reliability and scalability.

4.1 Human Evaluation Protocol

Human Evaluation Process. For LIMITGEN-Syn, we assess whether they correctly identify the intended subtype and calculate the accuracy. For LIMITGEN-Human, we assess the generated limitations across three dimensions: faithfulness, soundness, and importance. The detailed criteria are presented in the Appendix A.3.

For each criterion, Likert-scale scores ranging from 1 to 5 are used. Given the paper and a limitation generated by LLM, human evaluators are asked to assign scores for each dimension. Initially, ground truth references are not provided, minimizing potential bias from direct comparisons to the reference, as LLMs can generate limitations that are reasonable but not explicitly included in peer reviews.

After submitting their initial scores, evaluators are then provided with the reference and asked to adjust their scores if they identify any aspects that may have been overlooked.

Ensuring Reliable and Reproducible Human Evaluation. To ensure the reliability and reproducibility of our human evaluation, we develop a detailed assessment guideline, provided in Appendix A.3. To measure inter-annotator agreement, we sample 50 fixed generated instances from LIMITGEN-Syn and LIMITGEN-Human, each independently assessed by two expert annotators. In LIMITGEN-Syn, the resulting Cohen's Kappa score is 0.833. In LIMITGEN-Human, the scores for the criteria of importance, faithfulness, and soundness are 0.772, 0.735, and 0.717, respectively, indicating a high level of consistency among evaluators.

4.2 Automated Evaluation Protocol

To automatically evaluate the quality of the generated limitations, we compare them with the ground-truth limitations using a two-step process.

Coarse-grained Evaluation. For LIMITGEN-Syn, we use GPT-4o to classify the generated limitations and assess whether they correctly identify the intended subtype. Accuracy is used as the evaluation metric: a sample is deemed correct in the coarse-grained evaluation if at least one generated limitation accurately matches the subtype. For LIMITGEN-Human, we refer to MARG (D'Arcy et al., 2024), evaluating recall, precision, and Jaccard Index to measure the overlap between generated and ground truth limitations for a paper. These metrics are then averaged across all papers to produce a single aggregated value for each metric.

Fine-grained Evaluation. If a generated limitation correctly identifies the subtype or has a successful match in the ground truth limitations, we further evaluate the content to determine its alignment with the ground truth. This is achieved through reference-based evaluation using GPT-4o, which assigns scores to the generated limitations on from 1 to 5. These scores are based on two key criteria: relatedness to the ground truth and specificity in addressing the identified issue. Limitations that fail to determine the subtype or do not have a match during the coarse-grained evaluation are excluded from fine-grained evaluation and assigned a score of 0. For LIMITGEN-Syn, we calculate the average of the highest scores assigned to the limitations of each paper in fine-grained evaluation. For LIMITGEN-Human, we calculate the average of all limitations for each paper and then compute the overall average across all papers. This provides a holistic measure of the system's performance across both accuracy and quality dimensions.

Reliability Assessment. To validate the performance of our automated evaluation system, we also calculate the system correlation between the automated fine-grained evaluation and the human evaluation, using data presented in Table 3 and Ta-
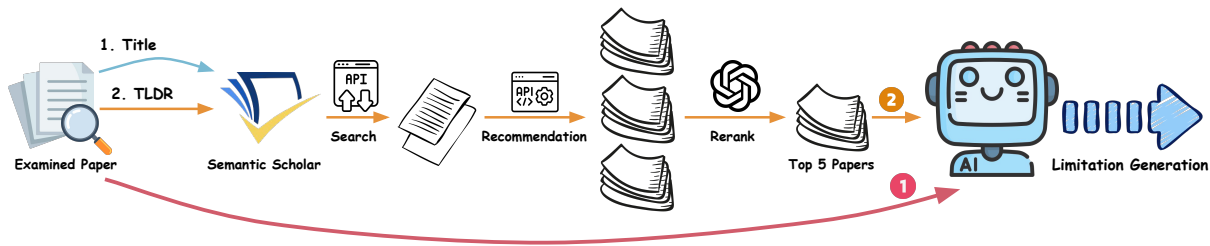
Figure 3: An overview of RAG pipeline. We prompt LLMs to query the Semantic Scholar API, retrieve recommended papers, and rerank them based on their abstracts.

ble 4. In LIMITGEN-Syn, the correlation between the fine-grained score and accuracy is 0.96. In LIMITGEN-Human, the correlation between the fine-grained score and faithfulness, soundness, and importance scores are 0.77, 0.60, and 0.67. By comparing with ground truth, our automated evaluation system can effectively assess the quality of the generated limitations.

5   Evaluated Systems

We next discuss the systems evaluated in our experiments, including LLMs, agent-based system, and RAG-enhanced pipeline.

5.1   Evaluated LLMs

We evaluate the performance of 4 frontier LLMs across two distinct categories in our benchmark: (1) Proprietary LLMs, including GPT-4o and GPT-4o-mini (OpenAI, 2024); and (2) Open-source LLMs, including Llama-3.3-70B (AI@Meta, 2024), Qwen2.5-72B (Yang et al., 2024). We require each model to generate the most significant limitations for an aspect of a paper. In the LIMIT-GEN-Syn experiments, we measure whether models identify the single most prominent limitation in each paper within their top three generated limitations, ensuring a fair comparison across systems.

5.2   Evaluated Agent-based System.

We also present our multi-agent approach for generating limitations. Our architecture, following MARG (D'Arcy et al., 2024), consists of a set of chat-based LLM agents (GPT-4o-mini in this study), each with its own chat history and prompt(s). The system includes three distinct agent roles: (1) a leader agent, responsible for coordinating tasks among agents; (2) a worker agent, which processes the full text of the paper; and (3) an expert agent, prompted to focus on a specialized

sub-task to support the leader. With task instructions for each aspect, the leader delegates specific sub-tasks to the other agents and synthesizes their responses to produce the final limitations.

5.3   RAG-Enhanced Limitation Generation

In preliminary testing, we observed that LLMs often fail to detect limitations or provide substantive suggestions due to a lack of knowledge in related areas. To address this, we enhanced the evaluated systems' capabilities by incorporating the RAG module, a method proven effective for knowledge-intensive tasks (Lewis et al., 2020; Shi et al., 2024), to ground limitation generation in the relevant literature. This method enables the LLMs to retrieve and consider related works when evaluating limitations in the given research paper.

Specifically, the retrieval process leverages the Semantic Scholar API and adapts based on the input paper's availability in the database. If the paper is available the database, we use its Semantic Scholar ID to fetch at most 20 recommended papers via the recommendation API[3]. If the paper is unavailable, we use GPT-4o-mini to generate a query based on the paper's abstract and use the relevance API[4] to identify related papers. From this search, the top 3 results are treated as seed papers, and for each seed paper, 5 additional recommendations are retrieved through the recommendation API, yielding a pool of 18 papers. These retrieved papers are then reranked by GPT-4o-mini, which assesses the similarity between the input paper and the candidates. The top 5 papers are selected.

Due to LLMs' context window constraints, directly providing all retrieved papers to them is impractical. We employ GPT-4o-mini to identify and

---

[3] https://api.semanticscholar.org/api-docs/recommendations
[4] https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_relevance_search

Systems	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy
Human	86.0%	3.52	82.0%
GPT-4o w/ RAG	52.0% +12.2%	1.34 +0.37	45.9% +16.0%
GPT-4o-mini w/RAG	49.1% +4.2%	1.25 +0.13	37.8% +5.9%
Llama-3.3-70B w/RAG	45.7% +2.4%	1.15 +0.05	32.7% +4.5%
Qwen-2.5-72B w/RAG	47.1% +1.2%	1.20 +0.03	31.5% +3.9%
MARG w/ RAG	68.1% +9.8%	1.83 +0.27	54.8% +17.7%

Table 3: Human and automated evaluation results of the LLMs and Agent-based system on LIMITGEN-Syn set averaged across all subtypes. For human evaluation, we randomly sample 100 examples from the dataset.

Systems	Automated Eval.		Human Eval. (1-5)		
	Jaccard	Fine.(0-5)	Faith.	Sound.	Import.
GPT-4o w/ RAG	15.9% +2.9%	0.42 +0.13	3.19 +0.49	2.84 +1.13	3.49 +0.60
GPT-4o-mini w/ RAG	15.5% +0.6%	0.39 +0.01	3.03 +0.28	2.78 +0.77	2.97 +0.53
Llama-3.3-70B w/ RAG	16.3% +0.1%	0.39 +0.04	2.98 +0.23	2.85 +0.70	3.05 +0.21
Qwen-2.5-72B w/ RAG	14.4% +1.0%	0.53 +0.11	2.91 +0.22	2.86 +0.35	2.94 +0.34
MARG w/ RAG	15.2% +2.5%	0.66 +0.24	3.60 +0.52	3.19 +0.98	3.78 +0.43

Table 4: Human and automated evaluation results of the LLMs and Agent-based system on LIMITGEN-Human set averaged across all aspects. We randomly sample 100 examples from the dataset for human evaluation.

extract content related to methodology, experimental design, result analysis, and literature review. This extracted content is then concatenated and used as a concise reference to help the LLMs effectively identify limitations in these aspects. In experiments involving MARG, we enable the expert agent to retrieve related papers and provide specific suggestions based on the retrieved content while refining the initial limitation comments.

6 Experiment Results

This section presents our main findings and in-depth analysis.

6.1 Results and Analysis

> 🛠 **RQ1:** How well do LLM-based systems perform in identifying limitations within scientific research?

Table 3 shows the performance of the evaluated systems on LIMITGEN-Syn. In Appendix B, we provide more detailed results on their performance for each subtype of limitation. The results demonstrate that identifying limitations in scientific papers remains a significant challenge for current LLMs. Even the best-performing LLM, GPT-4o, can only identify about half of the limitations that humans consider very obvious. Although MARG leverages multi-agent collaboration and generates more comments, successfully identifying more limitations, the feedback it provides still lacks specificity,

which is reflected in the fine-grained scores.

Table 4 shows the performance of the evaluated systems on LIMITGEN-Human, while the results on their performance for each aspect are illustrated in Appendix B. MARG outperforms all LLMs in terms of fine-grained scores and human evaluation but generates more comments than the other baselines, resulting in lower Jaccard scores. Consistent with the findings on LIMITGEN-Syn, the performance of all systems in LIMITGEN-Human remains quite poor. Their generated insights and feedback for top AI conference submissions lack depth and inspiration, especially when compared to those provided by experienced reviewers.

6.2 Analysis of RAG Pipeline

> 🛠 **RQ2:** Can RAG enhance LLMs' ability to identify limitations and provide constructive suggestions?

Overall Results. While LLMs currently struggle to identify limitations in scientific papers and provide constructive advice, there is potential for them to offer better feedback if they can retrieve relevant literature to address their gaps in domain knowledge and understanding of the research context. We conducted experiments on all evaluated systems with the integration of the RAG pipeline. As shown in Table 3 and Table 4, incorporating the RAG method can enhance LLM performance in refining their outputs.

Impact of Retrieved Content Quality on LLM Performance. We also investigate the impact of

Systems	Automated Eval.		Human Eval. (1-5)		
	Jaccard	Fine.(0-5)	Faith.	Sound.	Import.
GPT-4o-mini	15.0%	0.36	3.03	2.78	2.97
w/ RAG (Top 5)	+1.4%	+0.05	+0.28	+0.77	+0.53
w/ RAG (Top 3)	+1.3%	+0.04	+0.19	+0.56	+0.31
w/ RAG (Last 5)	+0.8%	+0.03	+0.07	+0.09	+0.05

Table 5: Human and automated evaluation results of different RAG settings from 18 retrieved papers on the subset of 100 examples from LIMITGEN-Human.

the quality of retrieved content on LLM performance. In LIMITGEN-Human, we randomly sample 100 examples and conduct experiments by GPT-4o-mini. For each example, we provide another two sets of retrieved papers as references: the top 3 ranked papers and the last 5 papers after re-ranking, from the 18 retrieved papers. The results, as shown in Table 5, demonstrate that providing a broader set of relevant papers, as in the standard RAG method with the top 5 papers, improves the LLM's performance in generating accurate limitations compared to using only the top 3 or the last 5 papers. RAG consistently provides some benefits, even when the retrieved papers are not the most relevant.

Case Study. We further conduct a case study to analyze the impact of RAG on LLM systems' ability to identify limitations. We select a total of 20 examples from both subsets, each successfully matching the targeted limitation subtype or receiving all three ratings of 4 or higher in human evaluation. See the Appendix B.4 for some of the examples. Retrieved external knowledge provides LLMs with up-to-date domain information and offers standard practices for addressing specific issues. By comparing relevant papers with the examined paper, LLM systems are better equipped to identify problems. Systems with stronger reasoning capabilities, such as GPT-4o and MARG, benefit the most from RAG, as they can leverage external information to derive meaningful insights and improve their analysis.

6.3 User Studies on Real-world Scenarios

> ✂ **RQ3:** How can this research be applied in real-world scenarios to assist human researchers in improving their work?

Our research focuses primarily on the AI domains. To investigate the applicability of our findings in more real-world scenarios, we design the following user studies to explore the domain generalization

User Study	Acc.
GPT-4o	
NLP Domain (as LIMITGEN-Syn)	45.9%
Biomedical Domain	31.3%
Computer Network Domain	37.5%
GPT-4o w/ RAG	
NLP Domain (as LIMITGEN-Syn)	61.9%
Biomedical Domain	50.0%
Computer Network Domain	56.3%
Llama-3.3-70B	
NLP Domain (as LIMITGEN-Syn)	32.7%
Biomedical Domain	25.0%
Computer Network Domain	31.3%
Llama-3.3-70B w /RAG	
NLP Domain (as LIMITGEN-Syn)	37.2%
Biomedical Domain	31.3%
Computer Network Domain	37.5%

Table 6: Human evaluation result of the adaptability of our research across different scientific domains.

of our research. Specifically, we examine the areas of biomedical sciences and computer networks. We first engage two experts in the two domains, each providing five research papers from their respective fields, focusing on those published after May 15, 2024, with which they are familiar. Following the annotation procedure outlined in LIMITGEN-Syn, the experts design perturbations across four aspects and annotate 32 examples in total. We then present another two experts with the perturbed papers and the generated limitations under two conditions: one utilizing our RAG pipeline and one without. As shown in Table 6, the human evaluation scores for GPT-4o and Llama-3.3-70B are consistent with the results observed in our main experiments. Our retrieval pipeline enhances the ability of LLMs to identify limitations. We believe that future work could further extend our research framework to encompass additional scientific domains.

7 Conclusion

This paper presents LIMITGEN, the first benchmark designed for systematically evaluating models on identifying and addressing scientific research limitations, supported by a reliable and systematic evaluation framework. We also demonstrate how RAG enhances limitation generation, showcasing its ability to help models identify weaknesses and provide more constructive feedback. Through a comprehensive analysis of LLM-based approaches for identifying different types of limitations, we offer key insights to guide future advancements.

Acknowledgments

Limitations

While our study provides valuable insights into the ability of LLM systems to identify limitations in scientific papers, several limitations remain that present opportunities for future work.

First, our work does not include non-textual inputs such as figures, which are integral to many scientific papers. As figures often provide crucial evidence or highlight key findings, future extensions to our benchmark could incorporate multimodal inputs to better evaluate LLMs' ability to identify limitations arising from inconsistencies or omissions in visual data.

Second, this study does not explore advanced RAG techniques. Our focus is on assessing the potential of LLM systems in this context rather than optimizing retrieval methods. We encourage researchers to build upon our benchmark and investigate advanced retrieval methods to further improve limitation identification.

Lastly, while our benchmark offers valuable insights into model performance, there are several limitations that should be considered. The current benchmark covers a limited time span, including some parts of 2024 and ICLR 2025, which may not fully represent the evolving landscape of research in the field. Given the rapid advancements in NLP, it is important to regularly update the benchmark to incorporate the latest publications. Another potential limitation lies in the reliance on our automated evaluation method. Inherent biases in these systems could affect the accuracy and reliability of the overall evaluation. Additionally, our taxonomy and benchmark focus primarily on AI, as this is the field we are most familiar with. Although we conducted a user study to assess its applicability to other domains, the nuances of different scientific disciplines may introduce challenges and limitation types that our framework does not fully address. Future work could expand this taxonomy by collaborating with experts from diverse fields, such as medicine, physics, and social sciences, to ensure broader generalizability.

Ethical Considerations

We have carefully considered the ethical implications of our work, which focuses on identifying limitations in scientific papers. Our approach is designed to assist human reviewers by offering complementary insights rather than replacing their essential role in the peer review process. We acknowledge potential risks, such as biases in LLM-generated outputs and the potential to undermine the integrity of scientific evaluations if these systems are misused. Our study emphasizes that LLMs are far from achieving the level of expertise and nuanced understanding of human experts. Future developments in this field should prioritize transparency, fairness, and risk mitigation to ensure these tools are employed responsibly. Furthermore, the raw paper data used in our study is collected from arXiv with distributed under the CC BY 4.0 (Creative Commons Attribution 4.0 International) license. In alignment with this licensing framework, we will release our dataset under the same CC BY 4.0 license. This ensures that our dataset remains freely accessible while requiring proper attribution to the original sources, thereby maintaining legal and ethical compliance with the terms under which the original data was shared.

References

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.

AI@Meta. 2024. The llama 3 herd of models.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15068–15083, Miami, Florida, USA. Association for Computational Linguistics.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing conversational dense retrieval via LLM-cognition data augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2700–2718, Bangkok, Thailand. Association for Computational Linguistics.

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

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Ryan Liu and Nihar B Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, et al. 2024. Aaar-1.0: Assessing ai's potential to assist research. *arXiv preprint arXiv:2410.22394*.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

OpenAI. 2024. Hello gpt-4o.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. CiteME: Can language models accurately cite scientific claims? In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

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. *arXiv preprint arXiv:2406.05688*.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. 2024. Scicode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. 2024. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. KIWI: A dataset of knowledge-intensive writing instructions for answering research questions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12969–12990, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Ren-Jing Cui, Chengcheng Han, Qiushi Sun, et al. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024a. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *International Conference on Language Resources and Evaluation*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024b. Hypothesis generation with large language models.

A LIMITGEN Benchmark

A.1 LIMITGEN-Syn

Table 8 illustrates the detailed distribution of the introduced limitation subtypes in our LIMITGEN-Syn.

A.2 LIMITGEN-Human

We randomly sample 1,000 papers from the ICLR 2025 submissions. We use GPT-4o to filter and classify the ground truth limitations, with the prompt provided in Figure 16 and Figure 15.

A.3 Annotator Guidelines

All annotators are experts with several NLP/ML publications as shown in Table 7. To ensure quality, they follow detailed annotation guidelines, which provide clear instructions for the annotation process.

Source Paper Collection   Our annotators follow the guidelines below to ensure only well-written arXiv papers are selected for perturbation:

- Exclude papers that do not focus on experimental work, such as surveys, position papers, and dissertations.

- Avoid papers with poorly written sections, lack of structure, or unprofessional presentation..

- Ensure the methods are well-defined, reproducible, and grounded in established scientific principles. Avoid papers with vague or unsupported claims.

- Select papers that provide thorough experiments, proper baselines, and detailed evaluations. The results should be well-documented and statistically sound.

- The paper should present a meaningful contribution to the field, such as a novel approach, insights, or applications, rather than incremental work.

Data Validation   When validating the perturbation, annotators should follow these guidelines carefully to evaluate whether the perturbations meet the intended quality standards:

- Check that the generated perturbation aligns with the limitation type specified in the instruction and verify that GPT-4o strictly follows the provided instruction to introduce the intended limitation.

- Ensure that all relevant sections needing modification are appropriately updated.

- Confirm that the perturbation does not compromise the clarity of the original text.

- Verify that the introduced limitation represents the most evident and significant limitation of the targeted aspect.

- Ensure the introduction of the limitation does not lead to unintended limitations elsewhere in the paper.

- The generated ground truth limitation should clearly articulate the problem and be reasonable.

Human Evaluation   For LIMITGEN-Human, we assess the generated limitations across three dimensions:

- **Faithfulness:** The generated limitations should accurately represent the paper's content and findings, avoiding any introduction of misinformation or contradictions to the original concepts, methodologies or results presented.

  - 5 points: Perfect alignment with the original content and findings, with no misinformation or contradictions. Fully reflects the paper's concepts, methodologies, and results accurately.
  - 4 points: Mostly aligns with the original content but contains minor inaccuracies or slight misinterpretations. These do not significantly affect the overall understanding of the paper's concepts or results.
  - 3 points: Generally aligns with the original content but includes several minor inaccuracies or contradictions. Some elements may not fully reflect the paper's concepts or results, though the overall understanding is mostly intact.
  - 2 points: Noticeable misalignment with the original content, with multiple inaccuracies or contradictions that could mislead readers. Some key aspects of the paper's concepts or results are misrepresented.
  - 1 point: Introduces significant misalignment by misrepresenting issues that do not exist in the paper. Creates considerable misinformation and contradictions that distort the original content, concepts, or results.

- **Soundness:** The generated limitations should be detailed and specific, with suggestions or critiques that are practical, logically coherent, and

ID	# NLP/AI Publication	Data Annotation	Data Validation	Human Evaluation	Human Performance
1	> 10	✓		✓	
2	> 10		✓		✓
3	5-10	✓	✓		
4	5-10	✓		✓	
5	1-5	✓			
6	1-5	✓			✓

Table 7: Details of annotators involved in dataset construction and LLM performance evaluation. LIMITGEN is annotated by experts in NLP domains, ensuring both the accuracy of the benchmark and the reliability of the human evaluation.

Property	Value
Methodology	250
# Low Data Quality	125
# Inappropriate Method	125
Experimental Design	250
# Insufficient Baseline	62
# Limited Datasets	63
# Inappropriate Datasets	63
# Lack of Ablation Study	62
Result Analysis	250
# Limited Analysis	125
# Insufficient Metrics	125
Experimental Design	250
# Limited Scope	83
# Irrelevant Citations	84
# Inaccurate Description	83

Table 8: Subtype distribution in the LIMITGEN-Syn subset.

purposeful. It should clearly address relevant aspects of the paper and offer insights that can genuinely improve the research.

- 5 points: Highly detailed and specific, with practical, logically coherent, and purposeful suggestions. Clearly addresses relevant aspects and offers insights that substantially improve the research.
- 4 points: Detailed and mostly specific, with generally practical and logically sound suggestions. Addresses relevant aspects well but may lack depth or novelty in some areas.
- 3 points: Detailed and specific but with some issues in practicality or logical coherence. Suggestions are somewhat relevant and offer partial improvements.
- 2 points: Somewhat vague or lacking in specificity, with suggestions that have limited practicality or logical coherence. Addresses relevant aspects only partially and provides minimal improvement.

- 1 point: Lacks detail and specificity, with impractical or incoherent suggestions. Fails to effectively address relevant aspects or offer constructive insights for improvement.

- **Importance:** The generated limitations should address the most significant issues that impact the paper's main findings and contributions. They should highlight key areas where improvements or further research are needed, emphasizing their potential to enhance the research's relevance and overall impact.

  - 5 points: Addresses critical issues that substantially impact the paper's findings and contributions. Clearly identifies major areas for significant improvement or further research, enhancing the research's relevance and overall impact.
  - 4 points: Identifies meaningful issues that contribute to refining the paper's findings and methodology. While the impact is notable, it does not reach the level of fundamentally shaping future research directions.
  - 3 points: Highlights important issues that offer some improvement to the current work but do not significantly impact future research directions. Provides useful insights for refining the paper but lacks broader implications for further study.
  - 2 points: Points out limitations with limited relevance to the paper's overall findings and contributions. Suggestions offer marginal improvements but fail to address more substantial gaps in the research.
  - 1 point: Focuses on trivial issues, such as minor errors or overly detailed aspects. Does not address substantive issues affecting the paper's findings or contributions, limiting its overall relevance and impact.

A.4 Human Baseline

To obtain an informative estimate of expert-level performance on LIMITGEN, we randomly sample 50 examples from each subset. Two expert annotators (*i.e.,* Annotators 1 and 6, as described in Table 7) independently solve these examples.

During human evaluation, the expert evaluators are not informed of the sources of these generated limitations. We report the evaluation results on Table 3 and Table 4.

A.5 Limitation Taxonomy

Aspect: Methodology
Limitation Subtype: Low Data Quality
Definition: The data collection method is unreliable, potentially introducing bias and lacking adequate preprocessing.
Paper Title: Leveraging Corpus Metadata to Detect Template-based Translation: An Exploratory Case Study of the Egyptian Arabic Wikipedia Edition
url: https://arxiv.org/html/2404.00565v1

Original Paper Content:
3.1 Dataset Filtrating and Labeling: We follow a few heuristic rules to classify Egyptian Wikipedia into articles created before and after the massive template-based translation activities related to creation dates ... since all articles are created by registered users and label the articles after translation as template-translated articles..
3.2 Dataset Preprocessing: We lightly preprocess the filtered articles by replacing all non-alphanumeric and non-Arabic characters with white spaces and normalizing the extra unnecessary whitespaces to one whitespace. We do not apply stemming, lemmatization, or any Arabic text normalization on the articles to have organic content (articles) as much as possible.
3.3 Dataset Filtrating and Labeling: We use two different types of embedding techniques to encode … The goal is to test with different embedding techniques to maximize the performance of our multivariate machine learning classifiers and investigate how the type and size of the word embeddings would affect their performance.

Perturbation Implementation:
3.1 Dataset Filtrating and Labeling: We follow a few heuristic rules to classify Egyptian Wikipedia into articles created before and after the massive template-based translation activities related to creation dates ... since all articles are created by registered users and label the articles after translation as template-translated articles..
~~3.2 Dataset Preprocessing: We lightly preprocess the filtered articles by replacing all non-alphanumeric and non-Arabic characters with white spaces and normalizing the extra unnecessary whitespaces to one whitespace. We do not apply stemming, lemmatization, or any Arabic text normalization on the articles to have organic content (articles) as much as possible.~~
3.3 Dataset Filtrating and Labeling: We use two different types of embedding techniques to encode … The goal is to test with different embedding techniques to maximize the performance of our multivariate machine learning classifiers and investigate how the type and size of the word embeddings would affect their performance.

Explanation: The lack of data preprocessing can result in the dataset containing non-alphanumeric and non-Arabic characters, which negatively impacts data quality.

Prompt: Read the following section of a scientific paper and remove any content related to data preprocessing methods, including {name of the method}. If the entire section discusses these topics, output 'None'.

Figure 4: An example of **Low Data Quality** and its perturbation implementation.
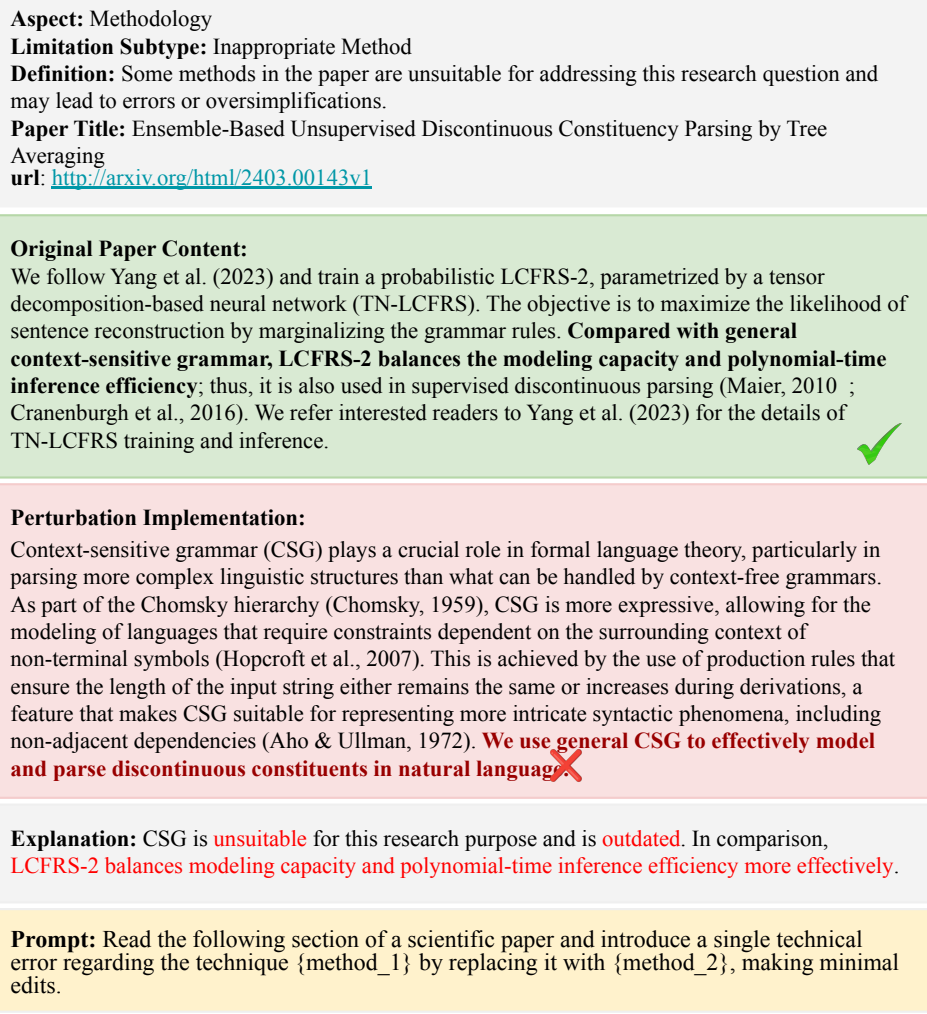
Aspect: Methodology
Limitation Subtype: Inappropriate Method
Definition: Some methods in the paper are unsuitable for addressing this research question and may lead to errors or oversimplifications.
Paper Title: Ensemble-Based Unsupervised Discontinuous Constituency Parsing by Tree Averaging
url: http://arxiv.org/html/2403.00143v1

Original Paper Content:
We follow Yang et al. (2023) and train a probabilistic LCFRS-2, parametrized by a tensor decomposition-based neural network (TN-LCFRS). The objective is to maximize the likelihood of sentence reconstruction by marginalizing the grammar rules. **Compared with general context-sensitive grammar, LCFRS-2 balances the modeling capacity and polynomial-time inference efficiency**; thus, it is also used in supervised discontinuous parsing (Maier, 2010 ; Cranenburgh et al., 2016). We refer interested readers to Yang et al. (2023) for the details of TN-LCFRS training and inference. ✔️

Perturbation Implementation:
Context-sensitive grammar (CSG) plays a crucial role in formal language theory, particularly in parsing more complex linguistic structures than what can be handled by context-free grammars. As part of the Chomsky hierarchy (Chomsky, 1959), CSG is more expressive, allowing for the modeling of languages that require constraints dependent on the surrounding context of non-terminal symbols (Hopcroft et al., 2007). This is achieved by the use of production rules that ensure the length of the input string either remains the same or increases during derivations, a feature that makes CSG suitable for representing more intricate syntactic phenomena, including non-adjacent dependencies (Aho & Ullman, 1972). **We use general CSG to effectively model and parse discontinuous constituents in natural language.** ❌

Explanation: CSG is unsuitable for this research purpose and is outdated. In comparison, LCFRS-2 balances modeling capacity and polynomial-time inference efficiency more effectively.

Prompt: Read the following section of a scientific paper and introduce a single technical error regarding the technique {method_1} by replacing it with {method_2}, making minimal edits.

Figure 5: An example of **Inappropriate Method** and its perturbation implementation.

Original Paper Content:
We compare UniMEEC with the baselines of MERC on IEMOCAP and MELD datasets, and the comparative results are shown in Table 2. Early works like BC-LSTM and DialogueRNN did not perform well on both datasets. Recent methods like MMGCN, and GA2MF achieve low performance in recognizing the happiness label for the IEMOCAP dataset and recognizing the disgust label for the MELD dataset. The low performance is caused by the dataset's label imbalance of emotion categories. Compared with the baselines, UniMEEC significantly improves WF1 by 1.99% and 1.85% on IEMOCAP and MELD datasets, respectively. Specifically, UniMEEC improves the emotion recognition performance on most emotion categories for two datasets. The possible reason for the improvements is that the unified framework of MERC and MECPE provides more auxiliary information, enhancing the interaction between emotion and emotion cause, thereby alleviating the label imbalance of IEMOCAP and MELD datasets. Furthermore, UniMEEC unifies the annotated labels of MERC and MECPE tasks and constructs a causal context between emotion and cause utterances, which implements the causality of response (emotion) and event (emotion cause). **In summary, UniMEEC consistently surpasses the state-of-the-art (SOTA) in most emotion category recognition on both datasets.** These results indicate the superiority of UniMEEC in emotion recognition and illustrate the effectiveness of a unified framework in model causality between MERC and MECPE. ✓

Perturbation Implementation: ✗
We present UniMEEC, which significantly improves WF1 by 1.99% and 1.85% on IEMOCAP and MELD datasets, respectively. Specifically, UniMEEC improves the emotion recognition performance on most emotion categories for the two datasets. The possible reason for the improvements is that the unified framework of MERC and MECPE provides more auxiliary information, enhancing the interaction between emotion and emotion cause, thereby alleviating the label imbalance of IEMOCAP and MELD datasets. Furthermore, UniMEEC unifies the annotated labels of MERC and MECPE tasks and constructs a causal context between emotion and cause utterances, which implements the causality of response (emotion) and event (emotion cause). These results illustrate the effectiveness of a unified framework in model causality between MERC and MECPE. ✗

Explanation: The revised section does not compare UniMEEC with the baseline models, making it difficult to demonstrate the effectiveness of the proposed method.

Prompt: Read the following section of a scientific paper and remove all content related to {baseline}, including the experiment on them, result analysis of them, and comparison to them.

Figure 6: An example of **Insufficient Baseline** and its perturbation implementation.

Aspect: Experimental Design
Limitation Subtype: Limited Datasets
Definition: Rely on limited datasets, which may hinder the generalizability and robustness of the proposed approach.
Paper Title: UniMEEC: Towards Unified Multimodal Emotion Recognition and Emotion Cause
url: https://arxiv.org/pdf/2404.00403v1

Original Paper Content:
We conduct experiments on four publicly available benchmark datasets of MERC and MECPE. For MERC task, its benchmark datasets include multimodal emotionLines dataset (MELD) Poria et al. (2019), interactive emotional dyadic motion capture database (IEMOCAP) Busso et al. (2008). IEMOCAP consists of 7532 samples, and each sample is labeled with six emotions for emotion recognition, including happiness, sadness, anger, neutral, excitement, and frustration. MELD contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, angry, surprise and disgust. For more details, please see Appendix A.For MECPE task, its benchmark datasets include **ConvECPE** Li et al. (2022a), and emotion-cause-in-friends (ECF) Wang et al. (2021). **ConvECPE is a multimodal emotion cause dataset constructed based on IEMOCAP, in which each non-neutral utterance is labeled with the emotion cause. It contains 151 dialogues with 7,433 utterances. Similarly,** Wang et al. (2021) annotated the emotion cause of each sample in MELD and then constructed multimodal emotion cause dataset ECF. ECF contains 1,344 conversations and 13,509 utterances. The detailed statistics of four datasets are shown in Table 1.For datasets IEMOCAP and MELD, we follow previous works Li et al. (2021c); Lu et al. (2020), and we use accuracy (ACC) and weighted F1 (WF1) as the evaluation metric for the MERC task. For datasets ECF **and ConvECPE**, we use precision (P), recall (R), and F1 as the evaluation metric for the MECPE task.

Perturbation Implementation:
We conduct experiments on four publicly available benchmark datasets of MERC and MECPE. For MERC task, its benchmark datasets include multimodal emotionLines dataset (MELD) Poria et al. (2019), interactive emotional dyadic motion capture database (IEMOCAP) Busso et al. (2008). IEMOCAP consists of 7532 samples, and each sample is labeled with six emotions for emotion recognition, including happiness, sadness, anger, neutral, excitement, and frustration. MELD contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, angry, surprise and disgust. For more details, please see Appendix A. For MECPE task, its benchmark dataset includes emotion-cause-in-friends (ECF) Wang et al. (2021). Wang et al. (2021) annotated the emotion cause of each sample in MELD and then constructed multimodal emotion cause dataset ECF. ECF contains 1,344 conversations and 13,509 utterances. The detailed statistics of the datasets are shown in Table 1. For datasets IEMOCAP and MELD, we follow previous works Li et al. (2021c); Lu et al. (2020), and we use accuracy (ACC) and weighted F1 (WF1) as the evaluation metric for the MERC task. For dataset ECF, we use precision (P), recall (R), and F1 as the evaluation metric for the MECPE task.

Explanation: The modified section only used a single dataset for the MECPE task, which may hinder the generalizability and robustness of the proposed approach.

Prompt: Read the following section of a scientific paper and remove all content related to {dataset}, including its introduction, experiments, and result analysis.

Figure 7: An example of **Limited Datasets** and its perturbation implementation.

Aspect: Experimental Design
Limitation Subtype: Inappropriate Datasets
Definition: Use of inappropriate datasets, which may not accurately reflect the target task or real-world scenarios.
Paper Title: UniMEEC: Towards Unified Multimodal Emotion Recognition and Emotion Cause
url: https://arxiv.org/pdf/2404.00403v1

Original Paper Content:
We conduct experiments on four publicly available benchmark datasets of MERC and MECPE. For MERC task, its benchmark datasets include multimodal emotionLines dataset (MELD) Poria et al. (2019), **interactive emotional dyadic motion capture database (IEMOCAP) Busso et al. (2008). IEMOCAP consists of 7532 samples, and each sample is labeled with six emotions for emotion recognition, including happiness, sadness, anger, neutral, excitement, and frustration.** MELD contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, angry, surprise and disgust. For more details, please see Appendix A.For MECPE task, its benchmark datasets include ConvECPE Li et al. (2022a), and emotion-cause-in-friends (ECF) Wang et al. (2021). ConvECPE is a multimodal emotion cause dataset constructed based on IEMOCAP, in which each non-neutral utterance is labeled with the emotion cause. It contains 151 dialogues with 7,433 utterances. Similarly, Wang et al. (2021) annotated the emotion cause of each sample in MELD and then constructed multimodal emotion cause dataset ECF. ECF contains 1,344 conversations and 13,509 utterances. The detailed statistics of four datasets are shown in Table 1.For datasets **IEMOCAP** and MELD, we follow previous works Li et al. (2021c); Lu et al. (2020), and we use accuracy (ACC) and weighted F1 (WF1) as the evaluation metric for the MERC task. For datasets ECF and ConvECPE, we use precision (P), recall (R), and F1 as the evaluation metric for the MECPE task.

Perturbation Implementation:
We conduct experiments on four publicly available benchmark datasets of MERC and MECPE. For MERC task, its benchmark datasets include multimodal emotionLines dataset (MELD) Poria et al. (2019), **Internet Movie Database (IMDB) Maas et al. (2011). IMDB consists of 50,000 samples, and each sample is labeled with binary sentiment categories, positive and negative.** MELD contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, angry, surprise and disgust. For more details, please see Appendix A.For MECPE task, its benchmark datasets include ConvECPE Li et al. (2022a), and emotion-cause-in-friends (ECF) Wang et al. (2021). ConvECPE is a multimodal emotion cause dataset constructed based on IEMOCAP, in which each non-neutral utterance is labeled with the emotion cause. It contains 151 dialogues with 7,433 utterances. Similarly, Wang et al. (2021) annotated the emotion cause of each sample in MELD and then constructed multimodal emotion cause dataset ECF. ECF contains 1,344 conversations and 13,509 utterances. The detailed statistics of four datasets are shown in Table 1.For datasets **IMDB** and MELD, we follow previous works Li et al. (2021c); Lu et al. (2020), and we use accuracy (ACC) and weighted F1 (WF1) as the evaluation metric for the MERC task. For datasets ECF and ConvECPE, we use precision (P), recall (R), and F1 as the evaluation metric for the MECPE task.

Explanation: The IMDB dataset does not contain multimodal data, making it unsuitable for the MERC task.

Prompt: Read the following section of a scientific paper. If {dataset_1} is mentioned, replace all content related to {dataset_1} with {dataset_2}, including substituting the introduction of {dataset_1} with the corresponding content for {dataset_2}. Do not modify unrelated content, such as the motivation, experimental setup, or methodology.

Figure 8: An example of **Inappropriate Datasets** and its perturbation implementation.

Aspect: Experimental Design
Limitation Subtype: Lack of Ablation Study
Definition: Fail to perform an ablation study, leaving the contribution of a certain component to the model's performance unclear.
Paper Title: Zero-shot and Few-shot Generation Strategies for Artificial Clinical Records
url: http://arxiv.org/pdf/2403.08664v2

Original Paper Content:
RQ1: Can a LLM achieve …?

RQ2: Does our proposed CoT prompting strategy improve the performance of prompt-based generation with LLMs?

Using our proposed CoT method provides improvement over our direct prompt in a zero-shot setting. Comparing the two approaches we see that using CoT prompting improves the performance of generation by 6.4 points, to the extent that the zero-shot Llama 2 model's performance with CoT prompting is comparable to, and slightly better than, the fine-tuned GPT-2 model. With regards to our research questions (RQs) we now answer RQ1 and RQ2. Firstly, we find that using our CoT prompting strategy, a zero-shot Llama 2 13b model loaded with 4-bit quantization can outperform a GPT-2 model fine tuned on EHR data in the same generation task, however it does not achieve the performance of more sophisticated fine-tuned models like Llama2 and BioGPT. In the case of the Llama 2 fine-tuned model this is to be expected, as it is the same model architecture trained on many examples for the same task. Secondly, and w.r.t RQ2, our CoT prompting method does improve zero-shot model performance compared to using a method that does not CoT.

RQ3: How do our prompting strategies perform in …?

Perturbation Implementation:
RQ1: Can a LLM achieve …?

RQ2: How do our prompting strategies perform in …?

Explanation: The lack of an ablation study makes it impossible to demonstrate the contribution of CoT to the performance improvement.

Prompt: Read the following section of a scientific paper and remove all content related to the ablation study {objective of the ablation study}, including the experiment process and result discussion. If the entire section discusses these topics, output 'None'.

Figure 9: An example of **Lack of Ablation Study** and its perturbation implementation.

Aspect: Result Analysis
Limitation Subtype: Limited Analysis
Definition: Offer insufficient insights into the model's behavior and failure cases.
Paper Title: MIPS at SemEval-2024 Task 3: Multimodal Emotion-Cause Pair Extraction in Conversations with Multimodal Language Models
url: http://arxiv.org/html/2404.00511v3

Original Paper Content:

3 Experiments

3.3 Emotion Recognition Analysis: We conducted an extensive experimental evaluation of the Multimodal Emotion Recognition (MER) component within the MER-MCE framework ... Consequently, models trained on these modalities exhibit subpar performance compared to the textual modality.

3.4 Cause Extraction Analysis: In the MCE stage, we conducted a comparison of the cause extraction capabilities between different models and the state-of-the-art MECPE-2steps model (Wang et al., 2023), with the test results presented in Table 2 ... However, as the number of windows increased further, the effectiveness gradually decreased due to the complexity of conversations with a larger historical context.

3.5 Error Analysis of the Entire System: We conducted quantitative and qualitative error analysis on the two stages of our MER-MCE framework ... It identified key areas for improvement, including handling facial occlusion, disambiguating emotional distractors, and capturing long-range dependencies in real-time settings. ✓

Perturbation Implementation:

3 Experiments

3.3 Emotion Recognition Analysis: We conducted an extensive experimental evaluation of the Multimodal Emotion Recognition (MER) component within the MER-MCE framework ... Consequently, models trained on these modalities exhibit subpar performance compared to the textual modality.

3.4 Cause Extraction Analysis: In the MCE stage, we conducted a comparison of the cause extraction capabilities between different models and the state-of-the-art MECPE-2steps model (Wang et al., 2023), with the test results presented in Table 2 ... However, as the number of windows increased further, the effectiveness gradually decreased due to the complexity of conversations with a larger historical context.

~~3.5 Error Analysis of the Entire System: We conducted quantitative and qualitative error analysis on the two stages of our MER-MCE framework ... It identified key areas for improvement, including handling facial occlusion, disambiguating emotional distractors, and capturing long-range dependencies in real-time settings.~~ ✗

Explanation: The absence of error analysis prevents a comprehensive understanding of the issues with the proposed approach.

Prompt: Read the following section of a scientific paper and remove all content related to the analysis of {topic}. If the entire section discusses these topics, output 'None'.

Figure 10: An example of **Limited Analysis** and its perturbation implementation.

Aspect: Result Analysis
Limitation Subtype: Insufficient Metrics
Definition: Rely on insufficient evaluation metrics, which may provide an incomplete assessment of the model's overall performance.
Paper Title: MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts
url: http://arxiv.org/html/2401.04081v2

Original Paper Content:
Accuracy and Perplexity - We observed that throughout the training of a variant of one of our smaller models, with 32 instead of 42 experts as presented in section 4.2, it maintains a lower perplexity than our strongest baseline (Transformer-MoE). **However, at the same time, Transformer-MoE consistently achieves higher accuracy than MoE-Mamba. We conjecture that this might be due to the fact that attention-based models are able to copy tokens verbatim, unlike SSM-based models, whose similar abilities might be hindered by the compression of the history into a finite hidden state.** We present **accuracy** and loss (log perplexity) plots alongside further discussion of those results in Appendix C.

Perturbation Implementation:
Perplexity - We observed that throughout the training of a variant of one of our smaller models, with 32 instead of 42 experts as presented in section 4.2, it **maintains a lower perplexity** than our strongest baseline (Transformer-MoE). We present loss (log perplexity) plots alongside further discussion of those results in Appendix C.

Explanation: The accuracy metric was not calculated, which led to the failure to discover that Transformer-MoE consistently achieves higher accuracy than MoE-Mamba, resulting in an incomplete assessment of the model's overall performance.

Prompt: Read the following section of a scientific paper and remove all content related to {metric}, including its introduction, experiments, and result analysis.

Figure 11: An example of **Insufficient Metrics** and its perturbation implementation.

Aspect: Literature Review
Limitation Subtype: Limited Scope
Definition: The review may focus on a very specific subset of literature or methods, leaving out important studies or novel perspectives.
Paper Title: MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts
url: http://arxiv.org/html/2401.04081v2

Original Paper Content:
2 Related Work
2.1 State Space Models and Related Attention-Free Architectures: State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling ... State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling.
2.2 Mixture of Experts: Mixture of Experts (MoE) is a class of techniques that allow drastically increasing the number of parameters of a model without much impact on the FLOPs required for the model's training and inference ... We also investigate the potential of integrating conditional computation into the Mamba block. ✔

Perturbation Implementation:
Related Work
2.1 State Space Models and Related Attention-Free Architectures: State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling ... State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling.
~~2.2 Mixture of Experts: Mixture of Experts (MoE) is a class of techniques that allow drastically increasing the number of parameters of a model without much impact on the FLOPs required for the model's training and inference ... We also investigate the potential of integrating conditional computation into the Mamba block.~~ ✘

Explanation: The modified section lacks an explanation of the current research and recent advancements in MoE, resulting in an incomplete related work section. This hinders the understanding of the method and leads to potential misinterpretations of its significance and applicability in the broader context.

Prompt: Read the following section of a scientific paper and remove all content related to the literature review of {topic}. If the entire section discusses these topics, output 'None'.

Figure 12: An example of **Limited Scope** and its perturbation implementation.

Aspect: Literature Review
Limitation Subtype: Irrelevant Citations
Definition: Include irrelevant references or outdated methods, which distracts from the main points and undermines the strength of conclusions.
Paper Title: MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts
url: http://arxiv.org/html/2401.04081v2

Original Paper Content:
2 Related Work
2.1 State Space Models and Related Attention-Free Architectures: State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling ... State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling.
2.2 Mixture of Experts: Mixture of Experts (MoE) is a class of techniques that allow drastically increasing the number of parameters of a model without much impact on the FLOPs required for the model's training and inference ... We also investigate the potential of integrating conditional computation into the Mamba block. ✔

Perturbation Implementation:
2 Related Work
2.1 State Space Models and Related Attention-Free Architectures: State Space Models …
2.2 Mixture of Experts: Mixture of Experts (MoE) …
2.3 Generative Adversarial Networks (GANs): GAN is introduced by Goodfellow et al. (2014), have revolutionized generative modeling by employing a dual-network structure: the generator, which creates data, and the discriminator, which evaluates its authenticity. Since their inception, numerous variants have emerged to enhance stability and output quality. For example, Conditional GANs (cGANs) allow for data generation based on additional input, while Deep Convolutional GANs (DCGANs) leverage convolutional architectures for improved image generation. Notable advancements include the Wasserstein GAN (WGAN) and its gradient penalty variant (WGAN-GP), which address training instability and mode collapse by utilizing a different loss function based on Wasserstein distance. GANs have found applications across diverse fields, including high-resolution image synthesis, medical imaging, and image-to-image translation (e.g., Pix2Pix). Despite their successes, challenges remain, such as mode collapse and evaluation difficulties, driving ongoing research into more robust architectures and training methodologies. ✘

Explanation: GAN is irrelevant to the topic of this paper. Discussing it in the related work section creates confusion and distracts from the main focus of the study, making it difficult for readers to grasp the relevance of the proposed approach and its contributions to the field.

Prompt: Read the following section of a scientific paper and write a short literature review in the same style about {an irrelevant topic}, approximately 50 words. Add this to the end of the section and output the entire revised text.

Figure 13: An example of **Irrelevant Citations** and its perturbation implementation.

Aspect: Literature Review
Limitation Subtype: Inaccurate Description
Definition: Provide an Inaccurate description of existing methods, which can hinder readers' understanding of the context and relevance of the proposed approach.
Paper Title: MoE-Mamba: Efficient Selective State Space Models with Mixture of Experts
url: http://arxiv.org/html/2401.04081v2

Original Paper Content:
State Space Models and Related Attention-Free Architectures - State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling. Stemming from signal processing, these models can be seen as a combination of **RNNs and CNNs (Gu & Dao, 2023).** Although they potentially offer considerable benefits, a number of issues have been identified with SSMs (Gu et al., 2022b), preventing SSMs from becoming the leading architecture in the task of **language modeling**. However, recent breakthroughs (Gu et al., 2022b; Fu et al., 2023; Smith et al., 2023; Gu & Dao, 2023), have allowed deep SSMs to be increasingly competitive against Transformers (Vaswani et al., 2017). In particular, Mamba (Gu & Dao, 2023), studied in this paper, has shown impressive results through its selective mechanism and hardware-aware design, which allows scaling to **billions** of parameters while retaining computational efficiency and strong performance. Besides SSMs, numerous other architectures have been proposed that do not rely on the quadratic attention mechanism (Zhai et al., 2021; Poli et al., 2023; Sun et al., 2023; Peng et al., 2023).

Perturbation Implementation:
State Space Models and Related Attention-Free Architectures - State Space Models (SSMs) (Gu et al., 2022b, 2021, a; Gupta et al., 2022; Li et al., 2022; Ma et al., 2022; Orvieto et al., 2023; Smith et al., 2023) form a family of architectures used for sequence modeling. Stemming from signal processing, these models can be seen as a combination of **Transformers (Vaswani et al., 2017)**. Although they potentially offer considerable benefits, a number of issues have been identified with SSMs (Gu et al., 2022b), preventing SSMs from becoming the leading architecture in the task of **computer vision**. However, recent breakthroughs (Gu et al., 2022b; Fu et al., 2023; Smith et al., 2023; Gu & Dao, 2023), have allowed deep SSMs to be increasingly competitive against Transformers (Vaswani et al., 2017). In particular, Mamba (Gu & Dao, 2023), studied in this paper, has shown impressive results through its selective mechanism and hardware-aware design, which allows scaling to **trillions of parameters** while retaining computational efficiency and strong performance. Besides SSMs, numerous other architectures have been proposed that do not rely on the quadratic attention mechanism (Zhai et al., 2021; Poli et al., 2023; Sun et al., 2023; Peng et al., 2023).

Explanation: The modified section contains factual errors regarding SSMs and Mamba, which affects readers' understanding and distorts the impact of this paper.

Prompt: Read the following section of a scientific paper and introduce a single factual error into the highlighted fact {original description} with minimal edits.

Figure 14: An example of **Inaccurate Description** and its perturbation implementation.

Limitation Aspect Classification

[System Input]:
Please classify the following limitation of a scientific paper into one of the following aspects: clarity, methodology, experimental design, result analysis, literature review, or others. Output only the corresponding aspect.

Classification Criteria:

Clarity: Issues in the presentation, structure, or language that hinder readers' understanding of the study's purpose, methods, results, or conclusions.
Methodology: Problems with the selection, application, or justification of research methods, such as unreliable data collection techniques or limited novelty, affecting the robustness or interpretability of findings.
Experimental Design: Shortcomings in the study's structure or execution, such as inappropriate dataset selection, lack of controls, or absence of ablation studies, which undermine the validity or generalizability of results.
Result Analysis: Problems in interpreting or presenting data, such as overgeneralization, missing case studies, or using inappropriate metrics, compromising the accuracy or validity of conclusions.
Literature Review: Issues with the literature review, such as omission of relevant studies, outdated sources, or incomplete synthesis of research, leading to biased conclusions or an incomplete understanding of the topic.

[User Input]:
Limitation:
{limitation}

Figure 15: Prompt for categorizing ground truth limitations in LimitGen-Human.

Limitation Filter

[System Input]:
Based on the review comments of a scientific paper below, retain only those comments that provide substantive suggestions. Merge comments discussing the same issue without omitting or summarizing any content. Present the final set of comments.

[User Input]:
Comments:
{comments}

Figure 16: Prompt for filtering ground truth limitations in LimitGen-Human.

B Experiments

B.1 Experiment Setup

```
Limitation Generation w/o RAG

[System Input]:
Read the following scientific paper and generate
major limitations in this paper about its {aspect}. Do
not include any limitation explicitly mentioned in the
paper itself and return only the limitations.


[User Input]:
Paper to review:
Title: {Title}
{Paper}
```

Figure 17: Prompt for limitation generation w/o RAG.

B.1.1 Agent-based System

We adopt the fundamental structure of MARG (D'Arcy et al., 2024), with modifications made better to align it with the requirements of our task. Each agent in the system begins with a unique "system" message at the start of its message history to provide specific instructions tailored to its role. For example, the "leader" agent is instructed to act as the leader; its role includes coordinating other agents to fulfill the user's requests. Additionally, the leader is guided to create a high-level plan based on its task instructions before initiating communication or delegating sub-tasks. Only the worker agent has access to the full content of the paper being reviewed. It is prompted to follow the leader's instructions to locate and summarize relevant content. The "expert" agent receives detailed instructions specific to their expertise area, focusing on particular subtasks they are responsible for. Both the leader and the expert agents can view the worker's responses, but the worker can only see the directives provided by the leader.

When a paper is input, the leader first organizes all agents to generate candidate initial comments collectively. Then, each comment is individually discussed and refined into a detailed limitation or discarded. In the RAG setting experiments, experts can reference related papers during the refinement stage. This enables them to leverage the latest literature to acquire domain-specific knowledge, thereby enhancing the quality and relevance of the generated feedback. We modified the prompts for

```
Limitation Generation w/ RAG

[System Input]:
Read the following content from several papers to
gain knowledge in the relevant field. Using this
knowledge, review a new scientific paper in this field.
Based on existing research, identify the limitations
of the 'Paper to Review'. Generate major limitations
in this paper about its {aspect}. Do not include any
limitation explicitly mentioned in the paper itself and
return only the limitations.


[User Input]:
Relevant Paper 1:
Title: {Title 1}
{Retrieved Content 1}


Relevant Paper 2:
Title: {Title 2}
{Retrieved Content 2}


Relevant Paper 3:
Title: {Title 3}
{Retrieved Content 3}

Relevant Paper 4:
Title: {Title 4}
{Retrieved Content 4}


Relevant Paper 5:
Title: {Title 5}
{Retrieved Content 5}


Paper to review:
Title: {Title}
{Paper}
```

Figure 18: Prompt for limitation generation w/ RAG.

each agent according to the specific requirements, which are presented from Figure 21 to Figure 33.

B.2 LIMITGEN-Syn Experiments

In this section, we discuss the detailed results for each subtype/aspect in LIMITGEN-Syn, as presented in Table 9 to Table 16

Overall, LLMs perform best in identifying limitations within the Result Analysis aspect of scientific papers. This may be due to the fact that this aspect often involves more directly interpretable and quantifiable elements, such as statistical results and performance metrics, which LLMs are well-equipped to assess. As a result, the integration of RAG provides minimal improvement in this aspect. In contrast, LLMs perform the weakest in identifying limitations within the Literature Review aspect,

Figure 19: Prompt for Coarse-grained Evaluation in
LIMITGEN-Syn.

as this aspect requires a deeper understanding of
the existing body of work and how it contextualizes
the paper being reviewed.

RAG demonstrates its greatest impact in the iden-
tification of limitations related to Experimental De-
sign. This is likely because referencing relevant
baseline methods or datasets from the retrieved pa-
pers helps enhance the specificity of the limitations.
By providing more concrete examples or compar-
isons, RAG enables LLMs to offer more detailed
and actionable suggestions, thereby improving the
overall quality of the generated limitations in this
area.

Within the same aspect, LLMs demonstrate vary-
ing abilities to identify different limitation sub-
types, and RAG also influences performance dif-
ferently across these subtypes. For instance, in the
Methodology aspect, the identification of limita-
tions related to low data quality outperforms that of
inappropriate methods. This discrepancy is likely
due to the inherent complexity of the inappropriate
method limitation, which requires a deeper under-
standing of the paper's core arguments and method-
ology. In contrast, low data quality limitations are
more straightforward and are often supported by
references from retrieved papers, which may in-
clude information on similar data preprocessing
techniques. As a result, RAG is particularly effec-
tive in assisting with the generation of limitations
related to low data quality.

Figure 20: Prompt for Fine-grained Evaluation in LIM-
ITGEN.

B.3 LIMITGEN-Human Experiments

Automated Evaluation Metrics Given a set of generated limitations C_{gen} and a set of ground truth limitations C_{gt} for a paper, each generated limitation is paired with every ground truth limitation of the same aspect. GPT-4o assesses the degree of relatedness for each pair, categorizing them as "none," "weak," "medium," or "high." Pairs rated "medium" or "high" are counted as successful matches. Using the alignments between C_{gen} and C_{gt}, we evaluate several metrics, as described below. We refer to MARG (D'Arcy et al., 2024) and define directional intersection operators $\overset{\leftarrow}{\cap}$ and $\overset{\rightarrow}{\cap}$ to represent the set of aligned elements in the left or right operand, respectively. For example, $C_{\text{gen}} \overset{\leftarrow}{\cap} C_{\text{gt}}$ is the set of elements of C_{gen} that align to any element in C_{gt}.

- **Recall:** $\frac{|C_{\text{gen}} \overset{\rightarrow}{\cap} C_{\text{gt}}|}{C_{\text{gt}}}$, the fraction of real reviewer comments that are aligned to any generated limitations.

- **Precision:** $\frac{|C_{\text{gen}} \overset{\leftarrow}{\cap} C_{\text{gt}}|}{C_{\text{gen}}}$, the fraction of generated limitations that are aligned to any ground truth limitation.

- **(Pseudo-)Jaccard:** The Jaccard index is a commonly used measure of set overlap. Let $Intersection = \frac{|C_{\text{gen}} \overset{\leftarrow}{\cap} C_{\text{gt}}| + |C_{\text{gen}} \overset{\rightarrow}{\cap} C_{\text{gt}}|}{2}$; then the Jaccard Index is $\frac{intersection}{|C_{\text{gen}}| + |C_{\text{gt}}| - intersection}$.

We adopt a macro-averaging approach at the individual paper level. We generate several limitations for each paper in the test set and compare them with the corresponding human-written limitations, calculating the relevant metrics for each comparison. These metrics are then averaged across all the papers to produce a single aggregated value for each metric.

Result Analysis Table 18 to Table 21 show the detailed result for all the aspects in LIMITGEN-Human.

Overall, LLMs exhibit higher overlap and better quality in generating limitations related to experimental design compared to human reviewers. This may be because experimental design often receives the most feedback from human reviewers, providing a clearer reference in the automated evaluation. Also, limitations in experimental design tend to be more structured and objective, which makes it easier for LLMs to identify and refine issues.

In contrast, LLMs perform the weakest in identifying limitations within the Literature Review aspect, which is consistent with the results observed in our LIMITGEN-Syn. And RAG proves to be most helpful in this aspect. By retrieving and incorporating relevant papers, RAG helps the model identify missing references, overlooked methodologies, or underexplored areas, leading to more comprehensive and informed limitations.

MARG Leader

You are part of a group that needs to perform tasks that involve a scientific paper. However, the paper is very long, so only the worker agent has it. You are the leader in charge of interacting with the user and coordinating the group to accomplish tasks. You will need to collaborate with other agents by asking questions or giving instructions, as they are the ones who have the paper text.
Communication protocol:
To broadcast a message other agents, write "SEND MESSAGE: " and then your message; alternatively, if you forget to include it until the end of your message, you can write "SEND FULL MESSAGE" and everything you just wrote will be sent. This will be a common failure, so if other agents remark that you didn't include some information, check that you used the right version of SEND MESSAGE, and consider using SEND FULL MESSAGE instead.

Additional instructions:
When you are given a task, your first step should be to draft a high-level plan with a list of steps, concisely describing how you will approach the task and your strategy for communicating with other agents. Then, execute the plan. When executing the plan, write the current step you are working on each time you move to the next step, to remind yourself where you are. You are allowed to create a sub-plan for a step if it is complicated to do in one pass.

You should continue to pay attention to details in the original task instructions even after you draft your plan. Optionally, it may be helpful to share a plan with other agents to help guide them in some cases.

Other agents do not know anything about the task being performed, so it is your responsibility to convey any information about the task that is necessary for them to provide helpful responses. You should make this part of your high-level plan. Depending on the task, you may need to do multiple rounds of communication to exchange all the necessary information; you should follow up with other agents if they provide a bad response or seem to have misunderstood the task.

In addition, depending on the responses you receive, you may need to ask follow-up questions, clarify your requests, or engage in additional discussion to fully reason about the task.

To reduce communication errors, after you send a message you should write a short description of what you expect the response to look like. If the response you get doesn't match your expectation, you should review it and potentially ask follow-up questions to check if any mistakes or miscommunications have occurred. It could be the case that an agent (including yourself) has misread something or made a logic error.

Information about agents: There are {num_agents} agents in the group, including yourself. You are {agent_name}. The other agent(s) are: {other_agent_names}.Write "Ready" if you have understood the assignment and the protocol to communicate with other agents. You will then be given tasks.

Figure 21: System prompt for the leader agent in MARG.

MARG Leader (Methodology)

[Initial Generation]

Task: Write a list of feedback comments, similar to the limitation section in a scientific paper. The main type of feedback you should focus on is the appropriateness and rigor of the methodology. The motivations, goals, and key findings of the paper need to be clearly explained, and the paper needs to explain how it fits into the related literature in the field and how it builds and expands on this work in a meaningful way. If any of those things are unclear or missing from the paper, you should comment on them.

Once you have established the paper's motivations, goals, and key findings, carefully scrutinize whether they are reasonable and well-justified or if they require additional procedures for improvement. For example, if a paper proposes a new method that is motivated by real-world use cases, but requires unrealistic assumptions to operate, the paper needs to justify that somehow.

Important: {expert} doesn't have a paper chunk, but they are good at coming up with questions and potential shortcomings of the paper's methodology. Explain the paper to {expert} and answer any questions they have with the help of the other agents until {expert} say they are finished. Write feedback based on any points {expert} indicates are in need of improvement.

Think carefully in a logical, step-by-step way. Ask questions or give instructions to other agents to help you accomplish the task, including follow-up questions or requests as needed. Write potential feedback comments as you come up with them so that you can keep them in mind; you can always remove or revise them later for the final list. It is helpful to show the final comments to other agents for confirmation, but if they do not respond after several rounds, there is no need to keep waiting.

Figure 22: Prompt for the leader agent in MARG on methodology.

MARG Leader (Experimental Design)

[Initial Generation]

Task: Write a list of feedback comments, similar to the comments in a peer-review. In addition, focus on major comments rather than minor comments; major comments are important things that affect the overall impact of the paper, whereas minor comments are small things like style/grammar or small details that don't matter much.

Be specific in your suggestions, including details about method or resource names and any particular steps the authors should follow. However, don't suggest things that have already been included or addressed in the paper. Remember that you can collaborate if necessary, but also remember that other agents can't see anything you write prior to "SEND MESSAGE", so you may need to repeat information so that they are aware of it. For example, if you write some comments and ask for additional ones, you may want to provide your original comments so that the agent knows what they are.

Your review comments should be specific and express an appropriate level of importance. For example, suppose a paper is missing some important baseline to show the contribution of the proposed method. A comment like "The authors could add more baselines, such as XYZ" is ineffective because it's too generic; even for a paper with strong experimental design, it's always possible to suggest additional baselines. This type of comment doesn't clarify if there's a significant issue with the current work. Instead, a more helpful comment would be: "The baselines in the paper are insufficient; key baselines, such as XYZ, should be considered. Without these, it's challenging to assess the contribution of the proposed method." Make sure your high-level plan mentions this instruction.

Some comments are a matter of degree. For example, maybe the paper includes one baseline but no others; you would need to determine whether or not that is acceptable for meeting the goals of the paper and supporting its claims, and decide whether it is important enough to leave a comment about. You can discuss with other agents as needed to help determine this.

You will need to communicate with other agents to understand the paper and learn what has already been addressed and what is still missing from the paper.

The main type of feedback you should focus on is the thoroughness of the experimental design. You should consider flaws in the design's robustness and ensure that potential biases are appropriately addressed. Your high-level plan should be roughly as follows:
1. Identify the research questions of the paper. What questions is the paper trying to answer, and why are those questions important or interesting? What findings does it contribute to the field? Ask the workers to go through the paper paragraph by paragraph and write down all the questions and claims, but do not include the experimental details.
2. Identify expectations for fulfilling the goals and claims. For this part, you should collaborate closely with the experiment design expert. Explain the task so that they can help you. Remember to put the information after SEND MESSAGE so that it gets sent correctly. Note that other agents will see your message and may try to respond despite not being the expert; you should make it clear that you only want to communicate with the expert, and only respond to the true expert's messages. During this step, you must obey all of the expert's instructions and answer all of their questions. The expert is {expert_1}.
a. Come up with a clear description of experiments including ablations that you would use to verify the paper's claims if you were doing the study yourself. Be specific and detailed in your description; what experiments should be conducted, how should they be set up, and why are they helpful for verifying the claims?
3. Check whether the paper matches your expectations
a. Go through the actual experiments in the paper. This will require communication with other worker agents to collect all the necessary information. Ensure that each aspect of the experiment is well-understood and documented to provide a comprehensive analysis of the paper's experimental design and results. This process should be detailed and cover all nuances needed for accurate evaluation and feedback. If agents do not provide all the needed information or if something is ambiguous, you must send additional messages to resolve the communication issues.
b. Let the expert to identify the similarities and differences between the actual experiment and your expected experiment. For each way the paper's experiments don't match your expectations, determine if this constitutes a shortcoming of the paper, or if the paper's experiments still fulfill the goals and claims of the paper. It may be helpful to share these comments with other agents to gather their opinions on whether the paper's experiments fall short.
b. If the paper's experiments are suboptimal or inadequate, write a feedback comment explaining the limitation and what the researchers should do next to resolve the issue. Be detailed and specific in your feedback to make it clear what the researchers in this field should do and why the suggestion is important. It is helpful to show the final comments to other agents for confirmation, but if they do not respond after several rounds, there is no need to keep waiting.

Figure 23: Prompt for the leader agent in MARG on experimental design.

MARG Leader (Result Analysis)

[Initial Generation]

Task: Write a list of feedback comments, similar to the limitation section in a scientific paper. The main type of feedback you should focus on is the analysis and presentation of the results. The motivations, goals, and key findings of the paper need to be clearly explained, and the paper needs to demonstrate how these results address the research questions. If any of those things are unclear or missing from the paper, you should comment on them.

Once you have established what the motivations, goals, and key findings of the paper are, you should check whether the results are analyzed and presented in a clear and effective manner. You should carefully scrutinize whether the statistical or analytical techniques are appropriate and correctly applied, and whether the results are presented with appropriate use of tables, figures, and charts. For example, if a paper presents results but fails to use appropriate statistical methods or if the presentation of results is unclear, the paper needs to address these issues.

Important: {expert_1} doesn't have a paper chunk, but they are good at coming up with questions and potential shortcomings of the paper's results analysis. Explain the paper to {expert_1} and answer any questions they have with the help of the other agents until {expert_1} say they are finished. You will likely need to pass their questions and comments along to the other agents that have the paper. Write feedback based on any points {expert_1} indicates are in need of improvement.

Think carefully in a logical, step-by-step way. Ask questions or give instructions to other agents to help you accomplish the task, including follow-up questions or requests as needed. Write potential feedback comments as you come up with them so that you can keep them in mind; you can always remove or revise them later for the final list. It is helpful to show the final comments to other agents for confirmation, but if they do not respond after several rounds, there is no need to keep waiting.

Figure 24: Prompt for the leader agent in MARG on result analysis.

MARG Leader (Literature Review)

[Initial Generation]

Task: Write a list of feedback comments, similar to the limitation section in a scientific paper. The main type of feedback you should focus on is the literature review. The paper's introduction, related work, and the overall context of the study need to be clearly evaluated. If any of those aspects are unclear or lacking, you should comment on them.

Once you have reviewed the introduction and related work sections, you should assess whether the literature review provides a comprehensive and relevant overview of existing research. Specifically, examine if the literature review is thorough and up-to-date, including recent and relevant studies. Additionally, check whether the literature review effectively contextualizes the current study and identifies gaps in existing research.

Important: {expert_1} doesn't have a paper chunk, but they are but they are skilled at identifying issues and gaps in literature reviews. Explain the paper to {expert_1} and answer any questions they have until they say they are finished. You will likely need to pass their questions and comments along to the other agents that have the paper. Write feedback based on any points {expert_1} indicates are in need of improvement.

Think carefully in a logical, step-by-step way. Ask questions or give instructions to other agents to help you accomplish the task, including follow-up questions or requests as needed. Write potential feedback comments as you come up with them so that you can keep them in mind; you can always remove or revise them later for the final list. It is helpful to show the final comments to other agents for confirmation, but if they do not respond after several rounds, there is no need to keep waiting.

Figure 25: Prompt for the leader agent in MARG on literature review.

MARG Leader (Refinement w/o RAG)

[Refine]

Refine and improve the following limitations that was written about a scientific paper. The goal is for the comment to be detailed and helpful, similar to a comment in the limitation section of a scientific paper. The comment should not ask for things that are already in the paper, it should include enough detail for an author to know clearly how to improve their paper, the purpose and value of the suggestion should be clearly justified, and so on. Remove the comment if it is bad (i.e., if it fails to meet those criteria). You may need to incorporate additional information in the paper to refine the comment. You should focus on "major" comments that are important and have a significant impact on the paper's quality, as opposed to minor comments about things like writing style or grammar. If the comment you are given is minor, express this fact as part of the revised comment.

Your revised review comment should be specific and express an appropriate level of importance. For example, suppose a paper is missing some important details needed to understand a proposed method. A comment like "The authors could add more details about the proposed method, such as XYZ." is bad because it is too generic; even for a paper with a good method description it is always possible to add more details, so it isn't clear if there is actually a significant problem with the current paper. Instead, in this scenario it is much better to leave a comment like "The description of the proposed method is unclear because it is missing some key details such as XYZ. Without these details it is hard to know whether ___.". Make sure your high-level plan references this instruction.

Note that only you are being given the comment; you will need to share it with other agents if you want them to have context. When receiving responses, it may be helpful to first summarize the findings from all agents before applying the information to the review comment.

Some comments are a matter of degree. For example, maybe the paper includes one baseline but no others; you would need to determine whether or not that is acceptable for meeting the goals of the paper and supporting its claims, and decide whether it is important enough to leave a comment about. You can discuss with other agents as needed to help determine this.

It may be helpful to work step-by-step examining one aspect of the comment at a time and considering what information is needed to verify that it is valid and important as well as what kind of clarification and rewording could help to make it clearer and more specific.

It is helpful to show the final comments to other agents for confirmation, but if they do not respond after several rounds, there is no need to keep waiting.

Here is the comment:
{review_comments}

Figure 26: Prompt for the leader agent in MARG at the refinement stage w/o RAG.

MARG Leader (Refinement w/ RAG)

[Refine]

Refine and improve the following limitations that was written about a scientific paper. The goal is for the comment to be detailed and helpful, similar to a comment in the peer-review. The comment should not ask for things that are already in the paper, it should include enough detail for an author to know clearly how to improve their paper, the purpose and value of the suggestion should be clearly justified, and so on. Remove the comment if it is bad (i.e., if it fails to meet those criteria). You need to understand the context of this comment with the help of other agents, and you may need to incorporate additional information from the paper to refine the comment. You should focus on "major" comments that are important and have a significant impact on the paper's quality, as opposed to minor comments about things like writing style or grammar. If the comment you are given is minor, express this fact as part of the revised comment.

You may also need to refer to the content from relevant papers. If you need to understand how other papers have addressed similar issues, you should collaborate closely with the experiment design expert. Give them information about the paper's topic and explain the task clearly so they can provide relevant references, such as additional datasets or recommendations for ablation study settings. The expert is {expert_2}. It's important to note that the relevant papers may not share the same goal as the paper under review, so the methods, data, and baselines from those papers may not be directly applicable. Therefore, you should check whether the content recommended by the expert agent is applicable to the current paper.

Your revised review comment should be specific and express an appropriate level of importance. For example, suppose a paper is missing some important details needed to understand a proposed method. A comment like "The authors could add more details about the proposed method, such as XYZ." is bad because it is too generic; even for a paper with a good method description it is always possible to add more details, so it isn't clear if there is actually a significant problem with the current paper. Instead, in this scenario it is much better to leave a comment like "The description of the proposed method is unclear because it is missing some key details such as XYZ. Without these details it is hard to know whether ___.". Make sure your high-level plan references this instruction.

Note that only you are being given the comment; you will need to share it with other agents if you want them to have context. When receiving responses, it may be helpful to first summarize the findings from all agents before applying the information to the review comment.

Some comments are a matter of degree. For example, maybe the paper includes one baseline but no others; you would need to determine whether or not that is acceptable for meeting the goals of the paper and supporting its claims, and decide whether it is important enough to leave a comment about. You can discuss with other agents as needed to help determine this.

It may be helpful to work step-by-step examining one aspect of the comment at a time and considering what information is needed to verify that it is valid and important as well as what kind of clarification and rewording could help to make it clearer and more specific.
Here is the comment:
{review_comments}

Figure 27: Prompt for the leader agent in MARG at the refinement stage w/ RAG.

MARG Worker

Figure 28: System prompt for the worker agent in MARG.

MARG Expert (Methodology)

[System Prompt]

You are part of a group of agents working with a scientific paper. You are highly curious and skeptical of papers, and your job is to help ensure that the paper has clearly explained the appropriateness and rigor of its methodology. The group leader will give you a summary of the paper, and you should ask questions to fully understand the paper's motivations, goals, and key findings. This includes asking follow-up questions as needed.

Scrutinize the paper heavily, focusing on identifying any hidden assumptions or potential issues that could undermine the appropriateness, motivation, contribution, and rigor of the methodology. For example, suppose a paper proposes a robot navigation algorithm that implicitly works only with omnidirectional instantly-accelerating robots; a questionable hidden assumption in this case would be that real-world robots can effectively be treated as omnidirectional, which is often untrue. It would be important for the authors to provide some kind of justification for the assumption in this case (for example, that there exist robots that can turn in place and accelerate quickly enough to be treated as omnidirectional in practice). Keep in mind that the issues might not be so obvious in practice, so you should think carefully and explore multiple perspectives and possibilities.

Think of the kinds of questions a scientific paper reviewer might ask, or what they might suggest is confusing or poorly justified in the paper.

Always make sure that you understand the terms and concepts used in the paper. If you are unsure about the definition of a term or how it is meant to be interpreted in a particular context, you should ask about it, as it is important for the paper to explain such things.

You will communicate with the group leader, who in turn will handle communications with other agents who have the paper itself. Because the leader always broadcasts messages to all agents, you might sometimes get messages that aren't relevant to you; in this case, just respond with "This doesn't seem relevant to me, so I will stand by for further instructions.". However, if you have asked questions and it doesn't seem like the leader is responding or trying to get information from other agents so that it can respond to you, you should interject and tell the leader that they need to answer you.

When you are done talking with the group leader, tell them that you are done with your review, and give them a summary list of any missing information, poorly justified points, or other suggestions that you identified.

Figure 29: Prompt for the expert agent in MARG on methodology.

MARG Expert (Experimental Design)

Figure 30: Prompt for the expert agent in MARG on experimental design.

MARG Expert (Result Analysis)

[System Prompt]

You are part of a group of agents evaluating a scientific paper. Your role is to ensure that the results are thoroughly analyzed and clearly presented. The group leader will provide a summary of the paper, and you should ask questions to fully understand how the results were obtained, their presentation, and their relevance to the research questions. This includes asking follow-up questions as needed.

Scrutinize the paper rigorously, focusing on evaluating the clarity and appropriateness of the results analysis. For example, consider whether the statistical or analytical techniques used are suitable and correctly applied. Check if the results are presented clearly with effective use of tables, figures, and charts. Assess whether the results directly address the research questions, and whether the findings are both statistically significant and practically meaningful. Ensure that the interpretation of the results is consistent with the data presented. Be thorough in your examination and explore multiple perspectives to identify any issues that could undermine the validity or clarity of the results analysis.

Think of the kinds of questions a scientific paper reviewer might ask, or what they might suggest is confusing or poorly justified in the paper.

Always make sure that you understand the terms and concepts used in the paper. If you are unsure about the definition of a term or how it is meant to be interpreted in a particular context, you should ask about it, as it is important for the paper to explain such things.

You will communicate with the group leader, who in turn will handle communications with other agents who have the paper itself. Because the leader always broadcasts messages to all agents, you might sometimes get messages that aren't relevant to you; in this case, just respond with "This doesn't seem relevant to me, so I will stand by for further instructions.". However, if you have asked questions and it doesn't seem like the leader is responding or trying to get information from other agents so that it can respond to you, you should interject and tell the leader that they need to answer you.

When you are done talking with the group leader, tell them that you are done with your review, and give them a summary list of any missing information, poorly justified points, or other suggestions that you identified.

Figure 31: Prompt for the expert agent in MARG on result analysis.

MARG Expert (Literature Review)

[System Prompt]

You are part of a group of agents working with a scientific paper. You are highly curious and skeptical of papers, and your job is to help ensure that the paper has clearly explained the scope and depth of its literature review. The group leader will give you a summary of the paper, and you should ask questions to fully understand the paper's context, relevance, and background. This includes asking follow-up questions as needed.

Scrutinize the paper heavily, focusing on identifying any shortcomings or potential issues in the comprehensiveness and relevance of the cited literature. For example, suppose a paper on natural language processing relies on outdated data from the 1990s while recent studies have significantly advanced the field; a notable shortcoming in this case would be the failure to incorporate more recent and relevant research. It would be important for the authors to provide some kind of justification for their choice of literature or explain why newer studies were not included. Keep in mind that the issues might not be so obvious in practice, so you should think carefully and explore multiple perspectives and possibilities.

Think of the kinds of questions a scientific paper reviewer might ask, or what they might suggest is confusing or poorly justified in the paper.

Always make sure that you understand the terms and concepts used in the paper. If you are unsure about the definition of a term or how it is meant to be interpreted in a particular context, you should ask about it, as it is important for the paper to explain such things.

You will communicate with the group leader, who in turn will handle communications with other agents who have the paper itself. Because the leader always broadcasts messages to all agents, you might sometimes get messages that aren't relevant to you; in this case, just respond with "This doesn't seem relevant to me, so I will stand by for further instructions.". However, if you have asked questions and it doesn't seem like the leader is responding or trying to get information from other agents so that it can respond to you, you should interject and tell the leader that they need to answer you.\n\nWhen you are done talking with the group leader, tell them that you are done with your review, and give them a summary list of any missing information, poorly justified points, or other suggestions that you identified.

Figure 32: Prompt for the expert agent in MARG on literature review.

MARG Expert (Refinement)

[Refine]

You are part of a group of agents working with a scientific paper. Your job is to help ensure that the paper is appropriate and rigorous. You will be provided with the methodology content from relevant papers for reference. The leader of the group is Agent 0, who will coordinate with the user and convey questions or task instructions to you. When the leader sends a message to you asking for assistance in judging the quality of a paper, you should help based solely on your knowledge of other relevant papers and avoid introducing any content that is not present in the paper chunks.

It's important to note that the relevant papers may not share the same goal as the paper under review, so the methods from those papers may not be directly applicable. However, you can still provide references for the leader to consider and offer sufficient explanation of the context.

Sometimes you will need more information in order to understand the task of the paper; in these cases, you should send a message to request this information from other agents. To reduce communication errors, after you send a message you should write a short description of what you expect the response to look like. If the response you get doesn't match your expectation, it is not necessarily wrong, but you should review it and potentially ask follow-up questions to ensure that no mistakes or miscommunications have occurred.

Be detailed and specific in the suggestions you give. What should the setup be? What settings or baseline methods should be compared? What metrics should be used? Make it clear which specific details are important and why, and which details are unimportant.

Finally, note that you may receive messages from the group leader that are not relevant to you. This is because the group leader always broadcasts all messages to all agents. If you get an irrelevant message, simply respond by saying "This doesn't seem relevant to me, so I will stand by for further instructions.".

Figure 33: Prompt for the expert agent in MARG at the refinement stage.

Systems	Low Data Quality			Inappropriate Method		
	Automated Eval.		Human Eval.	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy	Coarse	Fine (0-5)	Accuracy
GPT-4o	44.0%	1.00	38.5%	48.0%	1.24	33.3%
GPT-4o w/ RAG	+16.0%	+0.56	+15.3%	+4.0%	+0.00	+8.4%
GPT-4o-mini	61.6%	1.67	46.2%	56.0%	1.10	41.7%
GPT-4o-mini w/RAG	+1.6%	+0.07	+7.6%	+4.0%	+0.08	+8.3%
Llama-3.3-70B	56.8%	1.38	46.2%	54.8%	1.01	25.0%
Llama-3.3-70B w/RAG	+2.7%	+0.02	+0.0%	+3.4%	+0.15	+0.0%
Qwen-2.5-72B	45.5%	1.16	23.1%	43.1%	0.86	16.7%
Qwen-2.5-72B w/RAG	+4.7%	+0.13	+7.7%	+0.4%	-0.03	+8.3%
MARG	83.3%	2.17	69.2%	82.4%	1.71	75.0%
MARG w/ RAG	+12.2%	+0.28	+23.1%	-3.5%	-0.18	+0.0%

Table 9: Human and automated evaluation results on Methodology in LIMITGEN-Syn set.

Systems	Insufficient Baseline			Limited Datasets			Inappropriate Dataset			Lack of Ablation Study		
	Automated		Human	Automated		Human	Automated		Human	Automated		Human
	Coarse	Fine	Acc.	Coarse	Fine	Acc.	Coarse	Fine	Acc.	Coarse	Fine	Acc.
GPT-4o	41.7%	1.00	50.0%	58.3%	1.83	50.0%	75.0%	2.08	66.7%	16.7%	0.33	16.7%
w/ RAG	+25.0%	+0.50	+16.7%	+16.7%	+0.34	+16.7%	+16.7%	+0.92	+33.3%	+8.3%	+0.42	+16.6%
GPT-4o-mini	27.0%	0.62	33.3%	64.5%	1.94	33.3%	73.0%	1.71	66.7%	4.8%	0.11	0.0%
w/RAG	+3.2%	+0.16	+0.0%	+8.1%	+0.24	+16.7%	+4.8%	+0.27	+0.0%	+0.0%	+0.00	+0.0%
Llama-3.3-70B	29.8%	0.88	16.7%	66.1%	2.05	66.7%	35.5%	0.84	28.6%	9.7%	0.27	0.0%
w/RAG	+16.0%	+0.39	+16.6%	-2.8%	-0.12	+16.6%	+2.6%	+0.14	+14.3%	+5.1%	+0.09	+16.7%
Qwen-2.5-72B	44.4%	1.27	33.3%	70.5%	2.10	50.0%	52.5%	1.48	42.9%	21.0%	0.53	16.7%
w/RAG	+3.2%	+0.11	+0.0%	-27.9%	-0.90	+0.0%	+6.2%	+0.01	+0.0%	+9.6%	+0.31	+16.6%
MARG	41.7%	1.33	50.0%	38.5%	1.38	33.3%	15.4%	0.38	28.6%	58.3%	1.67	50.0%
w/ RAG	+16.6%	+0.34	+16.7%	+30.7%	+0.85	+33.4%	+30.8%	+0.77	+38.1%	+16.7%	+0.91	+33.3%

Table 10: Human and automated evaluation results on Experimental Design in LIMITGEN-Syn set.

Systems	Limited Analysis			Insufficient Metrics		
	Automated Eval.		Human Eval.	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy	Coarse	Fine (0-5)	Accuracy
GPT-4o	72.0%	1.88	84.6%	52.0%	1.36	50.0%
GPT-4o w/ RAG	+28.0%	+0.64	+15.4%	+8.0%	+0.44	+8.3%
GPT-4o-mini	46.4%	1.10	46.2%	50.4%	1.43	41.7%
GPT-4o-mini w/RAG	+15.2%	+0.39	+7.6%	+0.0%	+0.03	+0.0%
Llama-3.3-70B	59.0%	1.42	53.8%	47.9%	1.33	41.7%
Llama-3.3-70B w/RAG	+20.2%	+0.56	+7.7%	-6.0%	-0.18	-8.4%
Qwen-2.5-72B	71.0%	1.72	61.5%	47.2%	1.26	41.7%
Qwen-2.5-72B w/RAG	+21.0%	+0.67	+15.4%	-4.0%	-0.09	+0.0%
MARG	84.2%	2.58	84.6%	88.0%	2.32	75.0%
MARG w/ RAG	+6.7%	+0.10	+15.4%	+4.0%	+0.12	+8.3%

Table 11: Human and automated evaluation results on Result Analysis in LIMITGEN-Syn set.

Systems	Limited Scope			Irrelevant Citations			Inaccurate Description		
	Automated Eval.		Human Eval.	Automated Eval.		Human Eval.	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy	Coarse	Fine (0-5)	Accuracy	Coarse	Fine (0-5)	Accuracy
GPT-4o	100.0%	2.69	87.5%	0.0%	0.00	0.0%	56.2%	1.25	50.0%
GPT-4o w/ RAG	+0.0%	+0.06	+12.5%	+0.0%	+0.00	+11.1%	+12.6%	+0.31	+25.0%
GPT-4o-mini	100.0%	2.72	75.0%	0.0%	0.00	0.0%	41.0%	1.00	37.5%
GPT-4o-mini w/RAG	+0.0%	+0.03	+12.5%	+1.2%	+0.04	+0.0%	+6.0%	+0.12	+12.5%
Llama-3.3-70B	97.4%	2.73	75.0%	2.5%	0.02	7.7%	15.4%	0.32	12.5%
Llama-3.3-70B w/RAG	-15.9%	-0.69	+0.0%	-2.5%	-0.02	-7.7%	+0.6%	+0.08	+12.5%
Qwen-2.5-72B	90.0%	2.35	87.5%	0.0%	0.00	0.0%	23.5%	0.48	25.0%
Qwen-2.5-72B w/RAG	+6.3%	+0.16	+12.5%	+0.0%	+0.00	+0.0%	-4.5%	-0.04	+0.0%
MARG	100.0%	3.00	100.0%	37.5%	0.94	22.2%	57.1%	1.29	37.5%
MARG w/ RAG	+0.0%	+0.06	+0.0%	+9.6%	+0.24	+11.1%	+7.2%	+0.28	+25.0%

Table 12: Human and automated evaluation results on Literature Review in LIMITGEN-Syn set.

Query Generation

Generate tldr in 5 words: {Abstract}

Figure 34: Prompt for query generation.

Rerank

[System Input]:
Given the abstracts of {number} papers and the abstract of a reference paper, rank the papers in order of relevance to the reference paper. Output the top 5.

[User Input]:
Paper 1: {Title 1}
{Abstract 1}

Paper 2: {Title 2}
{Abstract 2}

Paper 3: {Title 3}
{Abstract 3}

...

Reference Paper: {Reference Paper Title}
{Reference Paper Title}

Figure 35: Prompt for reranking the retrieved papers and selecting the top 5.

Extract Relevant Content (Methodology)

[System Input]:
Concatenate all the content from the methodology sections of a paper.
Remove sentences that are irrelevant to the proposed methodology or models, and keep details about key components and innovations.

[User Input]:
{sections}

Figure 36: Prompt for extracting content relevant to methodology from a paper.

Extract Relevant Content (Experimental Design)

[System Input]:
Concatenate all the content from the experimental design sections of a paper.
Remove sentences that are irrelevant to the experiment setup, and keep details about the datasets, baselines, and main experimental, ablation studies.

[User Input]:
{sections}

Figure 37: Prompt for extracting content relevant to experimental design from a paper.

Extract Relevant Content (Result Analysis)

[System Input]:
Concatenate all the content from the result analysis sections of a paper.
Remove sentences that are irrelevant to the result analysis of the experiments, and keep details about the metrics, case study and how the paper presents the results.

[User Input]:
{sections}

Figure 38: Prompt for extracting content relevant to result analysis from a paper.

Extract Relevant Content (Literature Review)

[System Input]:
Concatenate all the content from the literature review sections of a paper.
Remove sentences that are irrelevant to the literature review, and keep details about the related works.

[User Input]:
{sections}

Figure 39: Prompt for extracting content relevant to result analysis from a paper.

Systems	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy
GPT-4o	46.0%	1.12	35.9%
GPT-4o w/ RAG	+10.0%	+0.28	+11.9%
GPT-4o-mini	58.8%	1.39	43.9%
GPT-4o-mini w/RAG	+2.8%	+0.07	+8.0%
Llama-3.3-70B	55.8%	1.20	35.6%
Llama-3.3-70B w/RAG	+3.1%	+0.08	+0.0%
Qwen-2.5-72B	44.3%	1.01	19.9%
Qwen-2.5-72B w/RAG	-2.1%	-0.08	+8.0%
MARG	82.8%	1.94	72.1%
MARG w/ RAG	+4.4%	+0.05	+11.5%

Table 13: Human and automated evaluation results on Methodology in LIMITGEN-Syn set.

Systems	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy
GPT-4o	47.9%	1.31	55.6%
GPT-4o w/ RAG	+16.7%	+0.54	+22.2%
GPT-4o-mini	42.3%	1.09	44.4%
GPT-4o-mini w/RAG	+4.0%	+0.17	+5.6%
Llama-3.3-70B	35.3%	1.01	37.3%
Llama-3.3-70B w/RAG	+5.2%	+0.12	+15.9%
Qwen-2.5-72B	47.1%	1.34	42.1%
Qwen-2.5-72B w/RAG	-2.2%	-0.12	+0.0%
MARG	38.5%	1.19	37.3%
MARG w/ RAG	+23.7%	+0.72	+29.4%

Table 14: Human and automated evaluation results on Experimental Design in LIMITGEN-Syn set.

Systems	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy
GPT-4o	62.0%	1.62	67.3%
GPT-4o w/ RAG	+18.0%	+0.54	+11.9%
GPT-4o-mini	48.4%	1.27	43.9%
GPT-4o-mini w/RAG	+7.6%	+0.21	+3.9%
Llama-3.3-70B	53.5%	1.38	47.8%
Llama-3.3-70B w/RAG	+7.1%	+0.19	-0.3%
Qwen-2.5-72B	59.1%	1.49	51.6%
Qwen-2.5-72B w/RAG	+8.5%	+0.29	+7.7%
MARG	86.1%	2.45	79.8%
MARG w/ RAG	+5.4%	+0.11	+11.9%

Table 15: Human and automated evaluation results on Result Analysis in LIMITGEN-Syn set.

Systems	Automated Eval.		Human Eval.
	Coarse	Fine (0-5)	Accuracy
GPT-4o	52.1%	1.31	25.0%
GPT-4o w/ RAG	+4.2%	+0.13	+18.1%
GPT-4o-mini	47.0%	1.24	18.8%
GPT-4o-mini w/RAG	+2.4%	+0.06	+6.2%
Llama-3.3-70B	38.4%	1.02	10.1%
Llama-3.3-70B w/RAG	-5.9%	-0.21	+2.4%
Qwen-2.5-72B	37.8%	0.94	12.5%
Qwen-2.5-72B w/RAG	+0.6%	+0.04	+0.0%
MARG	64.9%	1.74	29.9%
MARG w/ RAG	+5.6%	+0.20	+18.0%

Table 16: Human and automated evaluation results on Literature Review in LIMITGEN-Syn set.

Overlap Evaluation

[System Input]:
Compare the following pair of limitations: one generated and one from the ground truth.

Assess the degree of relatedness and the level of specificity of the generated limitation compared to the ground truth limitation. Start by providing a brief explanation for each category, and then assign a rating. Present your assessment in JSON format as follows:

{
"relatedness_reason": "< Provide a brief explanation of why you assessed the relatedness as you did>",
"relatedness": "<Choose one of the following options: 'none', 'weak', 'medium', 'high'>",
"specificity_reason": "<Provide a brief explanation of why you assessed the specificity as you did>",
"specificity": "<Choose one of the following options: 'less', 'same', 'more'>"
}

[User Input]:
Ground truth limitation:
{ground truth}
Generated limitation:
{generated limitation}

Figure 40: Prompt for measuring overlap in LIMITGEN-Human.

Systems	Coarse.		Fine.		Accuracy	
	Median	Variance	Median	Variance	Median	Variance
GPT-4o	52.0%	7.4%	1.25	0.60	50.0%	6.9%
GPT-4o w/ RAG	66.7%	9.3%	1.56	0.78	66.7%	8.4%
GPT-4o-mini	50.4%	8.5%	1.10	0.64	41.7%	5.3%
GPT-4o-mini w/ RAG	60.0%	9.0%	1.46	0.69	50.0%	6.7%
Llama-3.3-70B	47.9%	7.8%	1.01	0.64	28.6%	6.1%
Llama-3.3-70B w/ RAG	45.8%	7.1%	1.16	0.47	33.3%	6.3%
Qwen-2.5-72B	45.5%	6.5%	1.26	0.50	33.3%	5.9%
Qwen-2.5-72B w/ RAG	43.2%	7.9%	1.17	0.56	33.3%	7.3%
MARG	58.3%	7.2%	1.67	0.58	50.0%	6.5%
MARG w/ RAG	75.0%	3.6%	2.23	0.43	75.0%	3.8%

Table 17: The median and variance of the results across subtypes in LIMITGEN-Syn set.

Systems	Automated Evaluation				Human Evaluation (1-5)		
	Recall	Precision	Jaccard	Fine (0-5)	Faith.	Sound.	Import.
GPT-4o	46.7%	20.9%	17.1%	0.43	3.14	2.82	3.62
GPT-4o w/ RAG	+2.1%	+2.8%	+2.6%	+0.12	+0.66	+1.17	+0.40
GPT-4o-mini	42.2%	19.4%	16.5%	0.38	2.92	2.66	3.10
GPT-4o-mini w/ RAG	+5.0%	+1.3%	+1.3%	+0.01	+0.30	+1.20	+0.56
Llama-3.3-70B	56.7%	19.9%	17.3%	0.42	3.04	2.78	3.24
Llama-3.3-70B w/ RAG	+2.8%	-0.1%	-0.4%	+0.03	+0.28	+1.09	+0.20
Qwen-2.5-72B	23.9%	21.6%	14.2%	0.49	2.52	2.78	2.94
Qwen-2.5-72B w/ RAG	+5.2%	+0.9%	+1.0%	+0.11	+0.30	+0.95	+0.16
MARG	51.9%	14.9%	12.4%	0.64	3.21	2.67	3.56
MARG w/ RAG	+1.9%	-1.1%	-0.7%	+0.24	+0.97	+1.41	+0.33

Table 18: Human and automated evaluation results on Methodology in LIMITGEN-Human set.

Systems	Automated Evaluation				Human Evaluation (1-5)		
	Recall	Precision	Jaccard	Fine (0-5)	Faith.	Sound.	Import.
GPT-4o	61.4%	34.1%	28.0%	0.70	3.70	3.41	4.12
GPT-4o w/ RAG	+4.5%	+2.6%	+3.2%	+0.13	+0.68	+1.14	+0.47
GPT-4o-mini	56.3%	33.3%	27.6%	0.70	3.76	3.40	3.66
GPT-4o-mini w/ RAG	-1.4%	-1.5%	-1.6%	-0.03	+0.28	+0.77	-0.17
Llama-3.3-70B	66.7%	31.9%	27.6%	0.63	3.28	3.63	3.56
Llama-3.3-70B w/ RAG	+2.6%	+0.0%	+0.1%	+0.06	+0.19	+0.57	-0.21
Qwen-2.5-72B	36.1%	42.0%	24.5%	0.94	3.79	3.63	3.67
Qwen-2.5-72B w/ RAG	+0.4%	+1.0%	+1.5%	+0.12	+0.10	+0.19	+0.25
MARG	54.9%	22.3%	18.8%	0.92	3.56	3.54	3.89
MARG w/ RAG	+1.5%	+3.8%	+2.6%	+0.21	+0.49	+0.70	+0.34

Table 19: Human and automated evaluation results on Experiment Design in LIMITGEN-Human set.

Systems	Automated Evaluation				Human Evaluation (1-5)		
	Recall	Precision	Jaccard	Fine (0-5)	Faith.	Sound.	Import.
GPT-4o	45.1%	14.8%	12.6%	0.36	3.21	2.73	3.41
GPT-4o w/ RAG	-1.5%	+2.8%	+1.8%	+0.09	+0.05	+1.15	+0.40
GPT-4o-mini	41.1%	15.5%	12.7%	0.32	3.07	2.93	2.87
GPT-4o-mini w/ RAG	+3.3%	+2.0%	+1.5%	+0.05	+0.46	+0.00	+1.14
Llama-3.3-70B	49.1%	15.5%	12.9%	0.32	3.00	2.73	2.81
Llama-3.3-70B w/ RAG	+4.5%	+0.6%	+0.6%	+0.03	+0.38	+0.12	+0.51
Qwen-2.5-72B	27.7%	18.9%	12.9%	0.48	2.86	2.81	2.91
Qwen-2.5-72B w/ RAG	+1.6%	+2.2%	+1.2%	+0.13	+0.32	+0.14	+0.48
MARG	59.4%	12.5%	11.1%	0.45	3.80	2.84	3.95
MARG w/ RAG	+4.2%	+3.3%	+2.2%	+0.16	+0.42	+1.08	+0.19

Table 20: Human and automated evaluation results on Result Analysis in LIMITGEN-Human set.

Systems	Automated Evaluation				Human Evaluation (1-5)		
	Recall	Precision	Jaccard	Fine (0-5)	Faith.	Sound.	Import.
GPT-4o	31.8%	6.6%	5.8%	0.18	2.71	2.38	2.81
GPT-4o w/ RAG	+10.9%	+5.2%	+4.1%	+0.17	+0.56	+1.10	+1.15
GPT-4o-mini	21.1%	6.3%	5.2%	0.14	2.37	2.14	2.24
GPT-4o-mini w/ RAG	+7.5%	+1.0%	+1.1%	+0.05	+0.09	+1.11	+0.61
Llama-3.3-70B	39.5%	8.6%	7.5%	0.20	2.60	2.25	2.59
Llama-3.3-70B w/ RAG	+2.8%	+0.1%	+0.1%	+0.01	+0.07	+1.04	+0.34
Qwen-2.5-72B	16.9%	7.8%	6.0%	0.22	2.46	2.24	2.22
Qwen-2.5-72B w/ RAG	-0.6%	+1.3%	+0.2%	+0.06	+0.18	+0.11	+0.50
MARG	65.9%	22.1%	18.4%	0.62	3.84	3.71	3.72
MARG w/ RAG	+4.3%	+5.3%	+5.9%	+0.36	+0.19	+0.71	+0.84

Table 21: Human and automated evaluation results on Literature Review in LIMITGEN-Human set.

Systems	Automated Eval.		Human Eval. (1-5)		
	Jaccard	Fine.(0-5)	Faith.	Sound.	Import.
GPT-4o	14.9%	0.39	3.17	2.78	3.52
w/ RAG	+2.2%	+0.11	+0.37	+1.15	+0.47
GPT-4o-mini	14.6%	0.35	3.00	2.80	2.99
w/ RAG	+1.4%	+0.03	+0.37	+0.76	+0.58
Llama-3.3-70B	15.1%	0.37	3.02	2.75	3.02
w/ RAG	+0.1%	+0.04	+0.33	+0.83	+0.32
Qwen-2.5-72B	13.6%	0.48	2.69	2.79	2.93
w/ RAG	+1.1%	+0.12	+0.31	+0.55	+0.32
MARG	15.4%	0.63	3.68	3.19	3.81
w/ RAG	+1.9%	+0.30	+0.44	+0.97	+0.38

Table 22: The median of the results across aspects in LIMITGEN-Human set.

Systems	Automated Eval.		Human Eval. (1-5)		
	Jaccard	Fine.(0-5)	Faith.	Sound.	Import.
GPT-4o	0.9%	0.05	0.16	0.18	0.29
w/ RAG	0.8%	0.04	0.28	0.20	0.12
GPT-4o-mini	0.9%	0.06	0.33	0.28	0.34
w/ RAG	0.7%	0.04	0.44	0.32	0.24
Llama-3.3-70B	0.7%	0.03	0.08	0.33	0.19
w/ RAG	0.7%	0.04	0.13	0.36	0.05
Qwen-2.5-72B	0.6%	0.09	0.38	0.33	0.35
w/ RAG	0.7%	0.10	0.31	0.48	0.25
MARG	0.2%	0.04	0.08	0.26	0.03
w/ RAG	0.4%	0.05	0.01	0.05	0.08

Table 23: The variance of the results across aspects in LIMITGEN-Human set.

B.4 Case study

Case Study:

Examined Paper: Self-Augmented In-Context Learning for Unsupervised Word Translation
Limitation Subtype: Limited Datasets (Experimental Design)

We adopt ~~two~~ one standard BLI benchmarks: 1) languages from XLING iincluding German (de), English (en), French (fr), Italian (it), and Russian (ru), their combinations resulting in BLI directions; ~~2) lower-resource languages including Bulgarian (bg), Catalan (ca), and Hungarian (hu) from PanLex-BLI, which result in BLI directions.888The datasets are also used in many recent BLI works . For both benchmarks,~~ a test set of K WT pairs is provided for each BLI direction. We experiment with four open-source LLMs: LLaMA , LLaMA-2, LLaMA , and LLaMA-2, found that other families of LLMs, including mT5, mT0, mGPT and XGLM, underperform LLaMA; we thus skip these LLMs in our work.
Implementation Details and BLI Evaluation. As mentioned in \u00a72 , our hyper-parameter and template search are conducted on a single, randomly selected language pair, which is de-fr, following Li et al.

Ground Truth: The study's reliance on limited datasets may affect the generalizability and robustness of the proposed approach. Conducting experiments on lower-resource languages such as PanLex-BLI could improve both the evaluation and applicability of the findings.

Retrieved Content 1: In this section, we describe the details of our main experiment. As indicated in Figure 1, **we consider five languages (Nepali, Sinhala, Gujarati, Kazakh, Turkish)** as the target unsupervised language pairs with English. We leverage auxiliary parallel data from six higher-resource languages (Chinese, Russian, Arabic, Hindi, Telugu, Tamil) with English. The domains and counts for the datasets considered can be found in Table 1 and a more detailed discussion on the source of the data and the preprocessing steps can be found in the Appendix.

Retrieved Content 2: We perform the CLWE initialization on the embedding layer of the encoder and decoder of both objectives and analyze its effectiveness in improving the lexical-level representations on the low-resource **English-Mizo** pair during the pre-training.

GPT-4o Output: The experimental setup relies heavily on a single language pair (de-fr) for hyper-parameter and template search, which may not generalize well to other lower-resource languages such as Nepali, Sinhala, Gujarati, Kazakh, Turkish and Mizo.

Figure 41: An example of GPT-4o w/ RAG generated limitation in LIMITGEN-Syn.

Case Study:

Examined Paper: AlignGPT: Multi-modal Large Language Models with Adaptive Alignment Capability
Limitation Subtype: Inappropriate Method (Methodology)

We propose to ~~add alignment vectors to the inputs of MLLMs~~ use a fixed alignment score for all image-text pairs to enrich their alignment capabilities. These ~~alignment vectors~~ fixed alignment scores are positioned ahead of the image embeddings and text embeddings. In the subsequent sections, we will elaborate on the role of ~~the alignment vectors~~ fixed alignment scores and the process to acquire them.

Ground Truth: The method employed in this paper, which applies a fixed alignment score to all image-text pairs without accounting for their actual alignment levels, is not suitable for this task. Instead, incorporating alignment vectors into the inputs of MLLMs would be more appropriate.

Retrieved Content: Given that multiple image patches often correspond to single words, we propose to learn a grouping of image patches for every token in the caption. To achieve this, we use a sparse similarity metric between image patches and language tokens and **compute for each token a language-grouped vision embedding** as the weighted average of patches. The token and language-grouped vision embeddings are then contrasted through a fine-grained sequence-wise loss that only depends on individual samples and does not require other batch samples as negatives. This enables **more detailed information to be learned** in a computationally inexpensive manner. SPARC combines this fine-grained loss with a contrastive loss between global image and text embeddings to learn representations that simultaneously encode global and local information.

GPT-4o Output: The paper employs a fixed alignment score for all image-text pairs during pre-training, which simplifies the integration process but fails to account for varying alignment levels across different pairs. This could lead to inaccuracies in understanding specific image-text relationships, as observed in real-world scenarios where alignment levels are not uniform.

Figure 42: An example of GPT-4o w/ RAG generated limitation in LIMITGEN-Syn.

Case Study:

Examined Paper: Improved Training Technique for Latent Consistency Models
Limitation Subtype: Literature Review

Consistency models are a new family of generative models capable of producing high-quality samples in either a single step or multiple steps. Recently, consistency models have demonstrated impressive performance, achieving results on par with diffusion models in the pixel space. However, the success of scaling consistency training to large scale datasets, particularly for text-to-image and video generation tasks, is determined by performance in the latent space. In this work, we analyze the statistical differences between pixel and latent spaces, discovering that latent data often contains highly impulsive outliers, which **significantly degrade the performance of iCT** (Song & Dhariwal, 2023) in the latent space.

Ground Truth: As an empirical paper, the authors seem to **have compared only with iCTs** that reproduce in hidden spaces. In fact, there have been many improvements on the consistent model of lifting hidden spaces, such as [1, 2], with which authors should compare or discuss. [1] Hyper SD: Trajectory Segmented Consistency Model for Effective Image Synthesis [2] Trajectory consistency disruption. The authors' experiments are limited to some simple modal datasets such as the FFHQ, CELEBA-HQ datasets. Empirical evidence without multi-modal datasets weakens persuasion.

Retrieved Content: Latent Consistency Model (LCM) extends the Consistency Model to the latent space and leverages the guided consistency distillation technique to achieve impressive performance in accelerating text-to-image synthesis. However, we observed that LCM **struggles to generate images** with both clarity and detailed intricacy. Consequently, we introduce **Trajectory Consistency Distillation (TCD)**, which encompasses trajectory consistency function and strategic stochastic sampling.

GPT-4o Output: A thorough identification of existing gaps in the current research landscape is necessary. Many studies have enhanced the consistent model of latent consistency models. The authors should critically examine areas where these efforts have fallen short and suggest concrete future research directions based on these identified gaps. This would position their work within the broader context of the field and clarify its significance in advancing knowledge.

Figure 43: An example of GPT-4o w/ RAG generated limitation in LIMITGEN-Human.

Case Study:

Examined Paper: EgoQR: Efficient QR Code Reading in Egocentric Settings
Limitation Subtype: Result Analysis

4.3 COMPARATIVE ANALYSIS
We conduct comprehensive comparative QR code quality analysis on internally collected, egocentric data for our QR code stack and off-the-shelf QR code pipelines. The internally collected dataset consisted of approximately 528 ego-centric images, with each image having 1 or more QR codes of different size. We define a successful reading as QR code being successfully detected and decoded. The success rate is defined as number of successful readings divided by number of QR codes. Table 2 shows the results of our runtime v.s. several off-the-shelf QR scan pipeline. By leveraging a light-weight super resolution model, we further improve the capability to scan extremely small QR code and lift the overall absolute scan success rate by 2%. Overall at egocentric setting, our relative scan success rate is at least 34% higher than the best off-the-shelf solution. We presented a novel, lightweight architecture for reading QR codes in egocentric settings, specifically designed to run on wearable devices **with minimal battery and latency impacts**.

Ground Truth: There is limited analysis of computational efficiency and power consumption, despite claims of suitability for wearable devices. The paper should provide concrete benchmarks on actual hardware and a thorough exploration of memory usage and resource demands to truly assess its performance in practical applications.

Retrieved Content: Report indicates that there is a growing demand for smart wearables, but **the availability of energy is limiting the future development of smart wearables**. Although the current cycle life of lithium-ion batteries can meet the needs of wearables, future wearables are moving toward the need for miniaturization, multi-functionality, and lightness. Lithium-ion batteries have limited capacity, and some special areas of wearables are unable to recharge and replace the batteries regularly. At the same time, from the perspective of environmental protection, novel energy supply methods have become very important for wearables.

GPT-4o Output: The paper lacks a comparison of resource consumption (e.g., battery, processing time) across different devices, particularly wearables, which is crucial for determining practical feasibility in the intended use cases of wearable devices under real-world settings

Figure 44: An example of GPT-4o w/ RAG generated limitation in LIMITGEN-Human.