BAGELS: Benchmarking the Automated Generation and Extraction of Limitations from Scholarly Text

Ibrahim Al Azher†, Miftahul Jannat Mokarrama†, Zhishuai Guo†,
Sagnik Ray Choudhury‡, Hamed Alhoori†
†Northern Illinois University, DeKalb, IL, USA
‡University of North Texas, Denton, TX, USA
{iazher1, mmokarrama1, zguo, alhoori}@niu.edu,
sagnik.raychoudhury@unt.edu

Abstract

In scientific research, "limitations" refer to the shortcomings, constraints, or weaknesses of a study. A transparent reporting of such limitations can enhance the quality and reproducibility of research and improve public trust in science. However, authors often underreport limitations in their papers and rely on hedging strategies to meet editorial requirements at the expense of readers' clarity and confidence. This tendency, combined with the surge in scientific publications, has created a pressing need for automated approaches to extract and generate limitations from scholarly papers. To address this need, we present a full architecture for computational analysis of research limitations. Specifically, we (1) create a dataset of limitations from ACL, NeurIPS, and PeerJ papers by extracting them from the text and supplementing them with external reviews; (2) we propose methods to automatically generate limitations using a novel Retrieval Augmented Generation (RAG) technique; (3) we design a fine-grained evaluation framework for generated limitations, along with a meta-evaluation of these techniques. Code and datasets are available at: Code: https://github.com/IbrahimAlAzhar/BAGELS_Limitation_Gen Dataset: https://huggingface.co/datasets/IbrahimAlAzhar/limitation-generation-dataset-bagels

1 Introduction

In scientific articles, "limitations" refer to the inherent shortcomings, constraints, or weaknesses of a study that may influence its results or restrict the generalizability of its findings (Ross and Bibler Zaidi, 2019). Such limitations can arise from various aspects of the research process, including the methodology, theoretical framework, data collection, experimentation, and analysis (Ioannidis, 2007). Authors commonly acknowledge issues such as internal validity concerns, measurement errors, confounding factors, and the omission of important variables (Puhan et al., 2009).

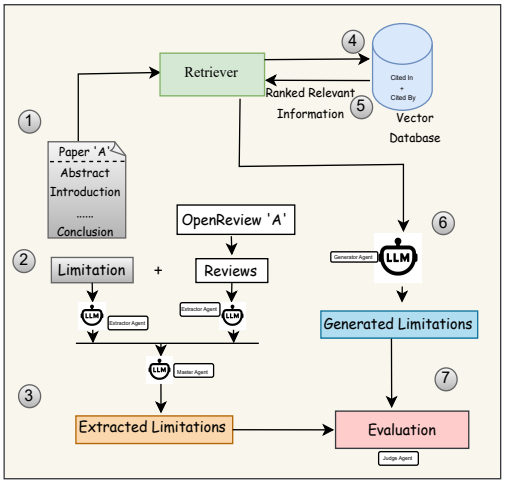

Figure 1: System architecture for dataset creation, limitation generation, and evaluation.

Openly discussing limitations is crucial. It upholds credibility and scientific integrity by demonstrating a commitment to ethical and transparent research practices (Bunniss and Kelly, 2010; Chasan-Taber, 2014; Annesley, 2010; Žydžiūnaitė, 2018). It also clarifies the scope of a study, supporting accurate interpretation, transferability, and reproducibility (Ioannidis, 2007; Eva and Lingard, 2008). In addition, it helps researchers avoid repeating the same shortcomings (Escande et al., 2016) while creating opportunities to refine methods and guide future research (Azher et al., 2025).

Despite these benefits, researchers are often reluctant to include limitations or articulate them in detail (Ioannidis, 2007; Ter Riet et al., 2013). Concerns about the potential impact on publication chances and career progression (Montori et al., 2004) can reinforce this tendency. Even when required to acknowledge limitations, as is now common in NLP/ML research, authors sometimes re-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19279–19294

sort to generic or irrelevant statements that obscure the study's real constraints (Ross and Bibler Zaidi, 2019). Moreover, limitations may serve as a form of *hedging*, where findings are presented cautiously to avoid making definitive claims (Hyland, 1998). This practice, while safer for authors, reduces the clarity and usefulness of the research.

Failure to disclose limitations undermines the scientific process and misleads readers, reviewers, and policymakers, preventing recognition of constrained findings and potential biases (Greener, 2018). Meanwhile, the volume of scientific publications has surged (Bornmann et al., 2021). These factors highlight the need for computational methods to study research limitations. However, progress in NLP toward automatic extraction, generation, and evaluation of limitations remains limited, largely due to the lack of standardized datasets, novel methods, and robust evaluation frameworks. This study takes a step toward closing this gap.

Our contributions are as follows (see Figure 1):

- **Dataset creation**. We build a dataset of research limitations by extracting them from papers and their reviews. By integrating author-reported and reviewer-identified limitations, this benchmark reduces self-reporting bias and provides a broader, more reliable resource for analyzing limitations and their impact on research.

- **Limitation generation**. We design a novel RAG system to automatically *generate* limitations, offering a way to supplement papers with high-quality, context-aware limitation statements.

- **Evaluation framework**. We introduce a new evaluation paradigm for generated limitations. Unlike traditional metrics (e.g., ROUGE (Lin, 2004), BLEU (Papineni, 2001), BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019)), which overemphasize common terms (e.g., bias, dataset, and generalizability), our framework leverages LLMs-as-judges for fine-grained, interpretable assessments and actionable error analysis.

2 Related Work

Several studies have examined how limitations are reported in papers. Ioannidis (2007) found that only 17% of top-tier articles mentioned limitations, with just 1% doing so in abstracts. Similarly, Puhan et al. (2012) reported that 27% of biomedical papers lacked limitations, risking overestimation of research reliability. Goodman et al. (1994) noted that acknowledging limitations is often problematic in peer review. Few journals require discussing limitations (Ioannidis, 2007), which can bias reviews and weaken scientific dialogue (Horton, 2002), highlighting the need for greater transparency.

Recent work has explored computational approaches to research limitations. Faizullah et al. (2024) proposed an LLM-chain pipeline to summarize and refine candidate limitations. Al Azher et al. (2024) integrated topic modeling with LLMs to derive structured limitation themes. Al Azher (2024) developed a graph-augmented LLM method for generating detailed limitation statements. Other studies address the shortcomings of visualizations by generating more meaningful captions for charts and graphs (Al Azher and Alhoori, 2024). However, these studies are limited to ACL/EMNLP corpora and rely on author-stated limitations, and use metrics such as ROUGE and BERTScore that miss finer-grained contextual alignment. Our framework addresses these shortcomings by including papers from different communities, incorporating Peer Review comments to capture reviewer perspectives, leveraging cited papers for broader context, and introducing a limitation-level evaluation method to preserve granularity.

Evaluating NLP outputs is essential for assessing quality, accuracy, and relevance. Traditional metrics like ROUGE and BLEU struggle with semantics, while BERTScore improves similarity but relies on references and lacks meaningful error analysis. Advances in large language models (LLMs) have opened new evaluation avenues (Zheng et al., 2023), from zero-shot and in-context learning (Wei et al., 2022) to specialized approaches such as GPTScore (Fu et al., 2023), TIGERScore (Jiang et al., 2023), and PandaLM (Wang et al., 2023). Other methods include AttrScore (Yue et al., 2023), which checks factual support, and SummacConv (Laban et al., 2022), which filters low-entailment sentences. Despite their promise, LLM-based evaluations face issues such as positioning bias, where input order can shift results. We address this by randomizing order and retaining stable outputs. More broadly, our evaluation advances beyond prior work by combining granularity-aware scoring, topic-level agreement, and LLMs-as-judges.

Taken together, prior research shows both the need and the opportunity for a more systematic treatment of research limitations. Building on these

insights, our work unifies dataset construction, limitation generation, and evaluation into a single framework, laying the foundation for more transparent and reproducible analysis of limitations.

3 Limitation Extraction & Evaluation

3.1 Dataset of Extracted Limitations

Granularity. A key challenge in building a dataset of research limitations is defining the appropriate level of granularity. Should a limitation be captured as a single phrase, a full sentence, or an entire paragraph? We define a **limitation** as a *sequence of sentences*, as individual sentences often do not encapsulate multiple limitations. In contrast, a single limitation can extend across multiple sentences, sometimes forming a complete paragraph.

Extraction Sources. Two primary sources form the basis of our dataset: (1) limitations explicitly acknowledged by authors, and (2) those highlighted through peer-review commentary. Although author-reported limitations often provide well-structured insights, previous research indicates that such limitations may be underreported or carefully hedged. To address this gap, we incorporate review comments, where reviewers often highlight additional constraints or weaknesses not mentioned by the authors.

Our dataset includes papers from major NLP and ML conferences, including ACL[1] and NeurIPS[2], as well as biomedical research from PeerJ [3]. We collect 6,932 NeurIPS papers (2021-2022), 5,739 ACL papers (2023-2024), and 1000 papers from PeerJ. In addition, we integrate OpenReview [4] comments for 2,802 papers from NeurIPS. All of the PeerJ papers contain self-reported limitations alongside other sections and peer review comments. For each paper, we use LLM to extract and get an average of 8 limitations from a paper and 10 from their reviews.

3.2 Extraction Process

We extract spans (blocks) of text from papers or review comments, and then refine them with LLMs, as opposed to passing in the entire paper to an LLM. This strikes a balance between accuracy and LLM usage cost.

1. Limitation *Span* Extraction: This step extracts blocks of text from the papers that correspond to limitations. We consider both explicit and implicit limitation statements:

a. Explicit limitations. These appear in a dedicated limitations section or subsection. We identify them using the AllenAI Science Parse tool [5], which segments papers into a structured JSON format, allowing for direct and reliable extraction of these dedicated sections. For peer review content in NeurIPS papers, we used Selenium to scrape the main review field from OpenReview, which typically includes both strengths and weaknesses of a paper.

b. Implicit limitations. These are embedded in broader sections such as discussion or conclusion. To identify them, we apply a Python regex script that searches for keywords such as *limitation(s)*, or *shortcoming(s)*. To improve precision, we exclude sections where limitations are rarely discussed (e.g., abstract, introduction, related work). Our script begins extraction when a limitation-related keyword is detected and continues until a terminal section marker is reached; extraction stops at terms such as acknowledgements, grant, future work, discussion, conclusion, or appendix. Although this process is effective, the regex approach for implicit limitations can occasionally capture irrelevant sentences, introducing noise into the results.

2. Refinement via LLM: To improve precision, we use an LLM to filter meaningful limitations from the tool-extracted ones (from both papers and review) by removing noisy sentences. Importantly, we strictly instruct the LLMs to extract limitation statements without paraphrasing, altering, or generating new content, and producing them as a structured sequence of sentences, denoted by $L_i = \{l_{i1}, l_{i2}, \ldots, l_{ix}\}$. To incorporate broader perspectives from peer reviews, we first aggregate comments from multiple review responses into a single consolidated text. We then prompt the LLMs (Figure 3, appendix) to segment this text and identify distinct limitation statements by reviewers, with the latter being denoted as $R_i = \{r_{i1}, r_{i2}, \ldots, r_{ix}\}$.

Following this extraction, a master LLM is tasked with merging the author-reported limitations L_i and the reviewer-identified limitations R_i of input paper P_i. The model is explicitly instructed to merge only those limitation statements that were

[1] https://aclrollingreview.org/cfp
[2] https://neurips.cc/public/guides/PaperChecklist
[3] https://peerj.com/benefits/indexing-and-impact-factor/
[4] https://openreview.net/
[5] https://github.com/allenai/science-parse

identical or semantically equivalent across both the author-mentioned limitations and the peer review. As before, the model is restricted from changing, rephrasing, or reordering any sentences during the merge process, and we get final *Ground truth extracted limitations* $G_i = \{g_{i1}, g_{i2}, \ldots, g_{ix}\}$. Finally, we evaluate the quality of these extracted and merged limitations through a user study described in § 3.3. We use GPT 4o-mini as both the extractor and master LLM.

3.3 Limitation Extraction Evaluation

Are the limitations extracted or generated? The first goal in the evaluation process is to check if the LLM extracted limitations are **grounded in the text**, i.e., they only come from the input (papers/reviews) and not from the LLMs' parametric knowledge or hallucinations. For this, we employ three annotators (separate from this paper's authors)[6].

The first ground truth consists of only author-mentioned limitations. We choose a sample of 100 limitations from ACL, NeurIPS, and PeerJ, and for each, we show them the source and ask a Yes/No question, whether they thought the LLM *extracted* the limitation from the source without generating text. Each annotator answer positively in $> 90\%$ of cases (avg $\pm$ std= $95 \pm 2.45\%$) (Table 1).

Model	Role	Sample	U1	U2	U3
GPT 4o-mini	Extractor	100	92	95	98

Table 1: Evaluating LLM as an extractor role with human annotator (U).

In the second evaluation, two annotators manually verified the extracted limitations from 1000 papers from NeurIPS and PeerJ *and their reviews*. The annotators assessed whether 1) each LLM-extracted author mentioned limitation was grounded in the source paper, (2) each extracted limitation from the peer review was also grounded in the review, and (3) the merged set (limitation + review) included only truly overlapping or matching limitations between the two sources. Their analysis confirmed that all extracted limitations were faithfully sourced, with no instances of hallucinated, noisy, or newly generated content. We also computed the performance of the Llama3 70B for this extraction task, and the result was unsatisfactory.

[6]CS graduate students with research experience in NLP and AI

The quality of the extraction. The SMEs from the last step annotated 500 ACL papers and 100 NeurIPS papers: one annotator extracted limitations (taking the full section when explicit, or selecting limitation-related sentences when implicit), and two others verified the results. We then compared the tool-based (GPT-4o mini) extractions against this gold standard. Notably, the human-extracted (gold) limitations were not segmented; therefore, we combined the LLM-extracted limitations and compared them with the gold ones using cosine similarity, precision, recall, F1, and fuzzy matching[7] (Table 2): ACL achieved a strong F1 of 85.69, likely aided by more frequent explicit limitation sections. NeurIPS yielded a moderate F1 of 72.42, reflecting the more scattered, implicit presentation of limitations where LLM should be utilized to remove noisy information.

Dataset	CS	P	R	F1	Fuzzy
ACL	89.38	89.63	84.93	85.69	91.18
Neurips	78.08	68.76	84.13	72.42	70.26

Table 2: Performance between Human Extracted Limitations vs Tool Extracted Limitations in Cosine Similarity (CS), Precision (P), Recall (R), F1 score (F1), and Fuzzy matching

3.4 Dataset Applications

The resulting dataset is publicly available[8] and can be used as a benchmark for evaluating automated limitation extraction and generation methods (§5). Beyond this, the extracted limitations can be examined and organized into a taxonomy of limitations in ML and NLP, offering a more structured understanding of common research challenges. By integrating this taxonomy into citation networks, we can introduce the concept of a *Limitation Multigraph*, enabling scientometric analyses into whether certain limitations shape the direction of subsequent research or, alternatively, tend to be overlooked. These avenues present new opportunities to study how the reporting (or the lack thereof) of limitations affects the broader scientific discourse, a topic we plan to explore in future work.

4 Limitation Generation

Most research papers either do not explicitly mention limitations or underreport them, even when a

[7]These strings are tokenized.

[8]https://huggingface.co/datasets/IbrahimAlAzhar/limitation-generation-dataset-bagels

dedicated section is provided. We compare two systems' ability to generate limitations from research papers: (a) vanilla LLM and (b) RAG. Note that the generators don't have access to the text from where the limitations are extracted, e.g., limitation sections of the papers, paragraphs identified as limitations, or paper reviews; otherwise, the task would be trivial. To improve computational efficiency, we use the three most important sections of a paper as input to the generators rather than the full text. The importance score is computed by the cosine similarity of a section and a reference limitation embedding (see Table 8, Appendix).

Vanilla LLM. In the *vanilla* LLM setup, when the input exceeds the context window, it is divided into chunks $\{P_i'\}$, and limitations are generated for each chunk (D'Arcy et al., 2024). The LLM is then also asked to aggregate these chunk-specific outputs into a cohesive, meaningful final set of limitations.

RAG Integration. A paper P_i can be used independently to generate limitations, but this approach risks overlooking valuable insights from other, potentially *related* papers. In particular, even when a paper lacks an explicit limitations section, other papers with similar methodologies or datasets may discuss relevant shortcomings. For example, a paper can use SVM and not explicitly mention the modeling assumptions, whereas a related paper possibly will. Moreover, certain findings may be implicitly contradicted by subsequent research. To address this issue, we employ a RAG framework, which allows the system to draw context from multiple papers rather than relying only on P_i.

There can be multiple notions of relatedness; we compare between two: a) relatedness induced by the citation network of P_i, and b) textual similarity between P_i and other papers. For the citation network, we use both the $\text{P}_{\text{cited-by, i}}$: papers citing P_i, and $\text{P}_{\text{cited-in, i}}$: papers that P_i cites. We parse the reference section of P_i to extract the DOI and title of each "cited in" work. The "cited by" DOI and titles are collected from the OpenAlex API [9]. We query the Semantic Scholar API [10] with P_i's title to get the DOIs for **top 5** most semantically close papers. These DOIs and titles are cross-referenced with arXiv metadata, and the full texts of the matched papers are downloaded and parsed with the Science Parse tool.

[9] https://openalex.org/

[10] https://www.semanticscholar.org/product/api

For each paper P_i, we build separate RAG indices with a) $\text{P}_{\text{cited-by, i}}$ b) $\text{P}_{\text{cited-in, i}}$, c) semantically close papers, and their combinations, where papers are split into chunks by section to preserve detail. We combine the strengths of both keyword-based (BM-25) and semantic (FAISS) search by assigning a 50% weight to the scores from each retriever. We use a *LLM-based reranker*, where we retrieve 20 chunks, and pass these chunks to a GPT 4o-mini model along with the original input paper. The model is prompted to score the relevance of each chunk on a scale of 1 to 10. Only the chunks that receive a relevance score of 8 or higher are ultimately selected. We compare this method with the simple baseline of just using the retrieved chunks in §7.4.

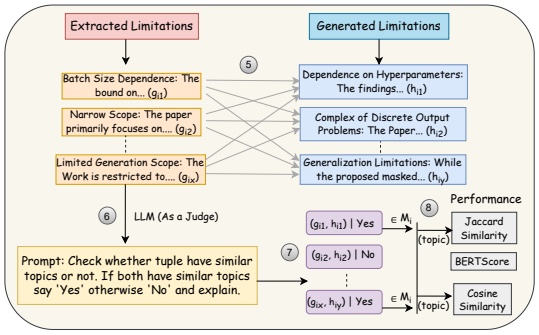

Figure 2: Evaluation of generated limitations.

5 Evaluation of Generated Limitations

We want to evaluate the quality of the generated limitations by comparing them with the *extracted* ones. Functionally, both the ground-truth and the predictions are a set of text segments. NLP metrics like BERTScore, ROUGE, and cosine similarity can yield surface overlaps, providing high scores even when the generated limitations are not appropriate, too generic, or imprecise. A possible alternative is to use a holistic LLM-as-Judge approach, where the generated/ground-truth limitations are merged into single text blocks and then compared. This lacks the point-level granularity needed for fine-grained analysis. We address both these problems by introducing the *PointWise* (**PW**) evaluation framework (Figure 2).

Problem Setup. Suppose we have a set of papers $P = \{P_1, P_2, \ldots, P_n\}$. For each paper P_i, we assume access to: *Ground truth limitations* $G_i = \{g_{i1}, g_{i2}, \ldots, g_{ix}\}$, where x is the number of ground truth limitations we extracted or annotated for P_i. And *LLM-generated limitations*

$H_i = \{h_{i1}, h_{i2}, \ldots, h_{iy}\}$, where y is the number of limitations produced by the LLM for P_i. Our goal is to measure (1) how many ground truth limitations the LLM correctly reproduces (*coverage*) and (2) how well each matched pair of limitations aligns in content and focus (*performance*).

5.1 Coverage

A. Pairwise Matching. To quantify coverage, we first create all possible pairs of limitations between the sets G_i and H_i. Let

$$S_i = \{(g_{ik},\, h_{il}) \mid 1 \le k \le x,\, 1 \le l \le y\}.$$

Hence, $|S_i| = x \times y$. We then use an LLM *as a judge* (Zheng et al., 2023) to decide if a ground truth limitation g_{ik} and a generated limitation h_{il} are similar in content or topic:

$$J(g_{ik}, h_{il}) \;=\; \begin{cases} 1, & \text{if } g_{ik} \text{ and } h_{il} \text{ are similar,} \\ 0, & \text{otherwise.} \end{cases}$$

We collect all *matched* pairs into a set

$$M_i \;=\; \{(g_{ik},\, h_{il}) \mid J(g_{ik},\, h_{il}) = 1\},$$

and let $|M_i| = z_i$ be the number of matched pairs for paper P_i.

B. Coverage of Ground Truth Limitations. We define $C_{Gi}(g_{ik}) = 1$ if the ground truth limitation g_{ik} appears in *at least one* matched pair in M_i, and 0 otherwise:

$$C_{Gi}(g_{ik}) \;=\; \begin{cases} 1, & \exists\, h_{il} \text{ such that } (g_{ik},\, h_{il}) \in M_i, \\ 0, & \text{otherwise.} \end{cases}$$

The *coverage of ground truth limitations* for paper P_i is

$$A_{Gi} \;=\; \frac{1}{x} \sum_{k=1}^{x} C_{Gi}(g_{ik}).$$

In other words, A_{Gi} measures the fraction of ground truth limitations in P_i that are matched with at least one LLM-generated limitation.

C. Coverage of LLM-Generated Limitations. Similarly, we define $C_{Hi}(h_{il}) = 1$ if a generated limitation h_{il} appears in *at least one* matched pair in M_i, and 0 otherwise:

$$C_{Hi}(h_{il}) \;=\; \begin{cases} 1, & \exists\, g_{ik} \text{ such that } (g_{ik},\, h_{il}) \in M_i, \\ 0, & \text{otherwise.} \end{cases}$$

The *coverage of LLM-generated limitations* for paper P_i is

$$A_{Hi} \;=\; \frac{1}{y} \sum_{l=1}^{y} C_{Hi}(h_{il}).$$

We aggregate these coverage values across all papers by taking their means:

$$A_G \;=\; \frac{1}{n} \sum_{i=1}^{n} A_{Gi}, \quad A_H \;=\; \frac{1}{n} \sum_{i=1}^{n} A_{Hi}.$$

D. Precision, Recall, and F_1. We also compute overall precision, recall, and F_1 scores. For each paper P_i:

$$\begin{aligned} \text{TP}_i &= |M_i|, \\ \text{FP}_i &= x - \sum_{k=1}^{x} C_{Gi}(g_{ik}), \\ \text{FN}_i &= y - \sum_{l=1}^{y} C_{Hi}(h_{il}). \end{aligned}$$

Here, TP_i (*true positives*) is the total number of matched pairs; FP_i (*false positives*) is the number of ground truth limitations not matched by any LLM-generated limitation; FN_i (*false negatives*) is the number of LLM-generated limitations unmatched by any ground truth limitation. True negative (TN_i) is not applicable in this case, as we do not have a defined *negative* class. If there is one ground truth limitation g_{ik} that matches with multiple LLM-generated limitations (and vice versa), True Positive (TP) counts as one. (Details in Appendix A.1)

5.2 Performance

After identifying matched pairs $(g_{ik}, h_{il}) \in M_i$, we score each pair's quality using (i) text-based metrics: ROUGE-L, BERTScore, and cosine similarity, and (ii) keyword overlap (Jaccard Similarity). Finally, the per-pair scores are averaged. Unmatched items are excluded, as our goal is to quantify similarity within aligned pairs rather than coverage (details in Appendix A.2).

6 Experimental Setup for Generation

We use three LLMs (GPT-3.5, GPT-4o-mini, and Llama 3.1 8B [11]) in a zero-shot setup for both the vanilla generation and RAG. The GPT models are

[11] https://huggingface.co/meta-llama/Llama-3.1-8B

accessed through APIs, and the LLama models are locally deployed with Ollama. For the vanilla generation, we also fine-tune three sequence-to-sequence models, T5 (512-token window) (Raffel et al., 2020), BART (1024 tokens), and Pegasus (1024 tokens) (Zhang et al., 2020) on a 70 / 30 train–test split. All models were trained for 3 epochs with a learning rate of 5×10^{-5}, weight decay 0.01, 300 warmup steps, and batch sizes of 4 (train) and 8 (eval), with early stopping; inputs longer than 512 tokens were truncated. For RAG, the vector database is built with llama-index [12], and the OpenAI text-embedding-ada-002 embedding model for encoding the source and query documents.

7 Experiments and Results

Evaluation of LLM as Aligner. The PointWise evaluation protocol above uses an LLM to determine whether a generated limitation matches or *aligns with* a ground truth one. We evaluate GPT 4o-mini's reliability in this task. A set of 100 positive (as per the model prediction) and 100 negative instances is annotated independently by two human evaluators. The human annotators have a Cohen's κ score of $> 92\%$ (Table 7, Appendix), which shows that the task is largely unambiguous. Cohen's κ between human annotators and model prediction is 90-95%, showing exceptional agreement. In comparison, Llama-3.1 400B shows poor agreement with the human judges, so in subsequent evaluations, we use GPT-4o mini as the aligner. See Table 9 in the appendix, for an example alignment.

7.1 Limitation Generation Evaluation

The ground truth contains papers that have a) only self-reported limitations and b) limitations coming from both self-reports and reviews.

7.2 Author-Mentioned Limitation

We evaluate the model's ability to generate self-reported limitations on the ACL part of the dataset, as these papers a) have explicit limitation sections, and b) do not have open-access reviews. The results are presented in Table 3.

Vanilla LLMs and fine-tuned models. Zero-shot models outperform trained models in almost all metrics, with GPT-3.5 achieving the best results in coverage metrics, and LLama 3 achieving the best in performance metrics. Surprisingly, GPT 4o-mini has a significantly worse performance than other zero-shot models. However, the performance metrics are based on n-gram overlap and embedding measures (e.g., ROUGE, BLEU, BERTScore, cosine similarity) that primarily capture surface overlap or shallow semantics and can miss factual correctness and completeness. Therefore, we prioritize coverage-based metrics, C_{GT}, C_{LLM}, and F1, and report NLP metrics as secondary diagnostics.

RAG. Since GPT-3.5 performs the best in the coverage metrics, we utilize it in a RAG setup, where the index consists of "cited-in" and "cited-by" papers. This improves the performance metrics, but comes at a cost of coverage metrics. To understand whether this reduction is caused by the RAG setup or the model, we include GPT 4o-mini in the same RAG setup, which shows a significant improvement in all metrics.

7.3 Self-reported & Peer-review Limitation

The ground truth here consists of author-stated limitations and peer-review limitations extracted from NeurIPS papers. We hypothesize that the RAG approaches should be beneficial for this dataset, as the reviewers are more likely to point out limitations from external sources, such as cited in/by or semantically similar papers. Therefore, we use this dataset to compare different RAG approaches with GPT 4o-mini as the baseline LLM, as the previous experiments (Table 3) suggest that it has the highest propensity of improvement with RAG.

Table 4 presents the performances with different RAG indices, with the first row representing vanilla LLM (no RAG). When the index is built with 100 random papers, the F1 score drops (-0.13) compared to the zero-shot approach. A combination of "cited in" and "cited by" papers achieves the highest F1 score of 0.67 – an increase of 0.02 over the baseline. However, when we further add the top five semantically related papers retrieved via the Semantic Scholar API, the F1 score reduces (-0.03), indicating that including loosely related content can introduce noise and reduce overall precision.

However, the performance metrics present a somewhat different story. Semantically related sentences from 100 randomly selected papers yield the highest scores across multiple metrics, including ROUGE-L, BERTScore, BLEU, and Jaccard similarity. We believe this is due to the vector database containing diverse texts, which are not semantically or n-gram overlapping with the ground truth. This perhaps also shows the brittleness of perfor-

[12] https://www.llamaindex.ai/

Model	Model type	R-L	BS	JS	CS	C_{GT}	C_{LLM}	Prec.	Recall	F1
T5	fine-tuned	19.92	87.81	10.82	31.79	35.48	29.59	0.29	0.31	0.30
BART	fine-tuned	19.43	87.67	10.68	31.91	33.71	30.10	0.30	0.31	0.31
Pegasus	fine-tuned	20.15	87.66	10.71	33.39	29.28	25.27	0.25	0.26	0.26
Llama 3	zero-shot	25.66	88.30	14.69	40.4	61.38	39.04	0.39	0.50	0.44
GPT-3.5	zero-shot	24.24	87.08	14.65	43.12	**76.62**	**46.65**	0.47	**0.67**	**0.55**
GPT 4o-mini	zero-shot	16.57	86.02	8.70	32.29	57.65	19.76	0.20	0.31	0.24
GPT-3.5 + RAG	zero-shot + RAG	**30.21**	**90.88**	**19.47**	**45.37**	39.99	44.66	0.42	0.40	0.41
GPT 4o-mini + RAG	zero-shot + RAG	23.17	87.33	12.99	39.29	67.13	45.67	**0.57**	0.45	0.51

Table 3: Results of models in "Coverage" (Coverage of Ground Truth Limitation (C_{GT}), LLM Generated Limitation (C_{LLM}), Precision, Recall, and F1-score) and "performance" metrics – **R**ouge-x, BLEU, BertScore (BS), Jaccard (JS) and Cosine (CS) similarity on the **ACL** dataset. In all metrics, a higher score denotes a better performance.

RAG Index	C_GT	C_LLM	F1	R-L	BS	CS	JS
Not applicable	67.34	63.81	0.65	15.30	86.66	33.43	8.69
100 Random Papers	60.31	48.87	0.52	**16.39**	**86.88**	32.34	**8.99**
Cited In	**69.27**	62.59	0.65	14.07	86.37	33.23	8.05
Cited By	68.45	64.01	0.65	14.49	86.34	**34.09**	8.36
Cited In + Cited By	68.84	**64.87**	**0.68**	14.35	86.35	33.94	8.28
Cited In + By + Semantically Similar 5 Papers	68.02	63.38	0.65	14.59	86.39	33.72	8.38

Table 4: Coverage evaluation of multiple types of RAG vector database settings in **NeurIPS 21-22 dataset** with GPT 4o-mini as the base LLM.

mance metrics for evaluating the generation quality of limitations.

7.4 Ablation Study

a. Size of the input text: We investigate the effect of the length of the input text on the generator models with an ablation study (Table 5). We use a) GPT-4o mini + RAG and b) Llama-3.1-8B, as these are the best-performing systems in the RAG and vanilla LLM setups, considering the author-mentioned limitations, the review-mentioned ones, and their combinations. When using GPT-4o-mini with RAG, expanding the context from the top-3 sections to all available sections generally increased pointwise scores for the author-written ground truth: C_{GT} (+0.94), C_{LLM} (+0.82), and F1 (+0.02). A similar trend was observed for the reviewer-suggested and combined ground truths, with the exception of a slight dip in the C_{GT} score for the combined (Auth + Rev) case. By contrast, most NLP-based metrics (e.g., ROUGE, BERTScore, and cosine) slightly decreased with all-section inputs. Taken together, this indicates that using only the top-3 sections is a cost-effective alternative in this setup: minor drops in pointwise metrics, small gains (or less drop) in NLP metrics, and no large performance loss overall. In Llama-3.1-8B, however, we observe the opposite trend. Moving to all sections produces a large F1 gain (+0.13) for the combined ground truth and improves most NLP-based metrics. This suggests the smaller Llama-3.1-8B benefits from the full-paper context to generate higher-quality limitations, whereas truncating to the top-3 sections leaves it under-informed.

b. Retriever Method: On the NeurIPS dataset, we evaluate our LLM re-ranker against a vanilla retriever baseline, both operating within a RAG framework with GPT-4o mini generator. While the baseline simply retrieves the top 3 chunks using a FAISS+BM25 search, our method re-ranks the top 20 chunks, leading to substantial gains in C_{GT} (+28.5), C_{LLM} (+15.53), and the F1 score (+0.24) (Table 6).

Our findings demonstrate that a multi-faceted approach, combining curated external data with targeted retrieval, significantly enhances the generation of scientific limitations. This is especially evident when we use limitations extracted from reviews in the ground truth, as the use of "cited in" and "cited by" papers in the RAG index achieves the highest F1 score. We also observe that the length of the input to the generator model has a different effect in the vanilla LLM and RAG setup. It might be beneficial to use full paper texts for smaller models, but larger models in RAG setups can perform reasonably well with the most important parts of a paper.

Metric	Input	Ground Truth		
	Sec	Auth	Rev.	Auth + Rev
GPT 4o-mini + RAG				
R-L	3	16.93	11.92	14.44
	All	16.73 (↓)	12.12 (↑)	14.35 (↓)
BS	3	87.27	86.26	86.43
	All	87.15 (↓)	86.21 (↓)	86.35 (↓)
CS	3	36.03	29.16	33.88
	All	35.52 (↓)	29.81 (↑)	33.94 (↑)
C_{GT}	3	82.60	61.19	69.78
	All	83.54 (↑)	61.93 (↑)	68.84 (↓)
C_{LLM}	3	29.83	62.69	61.59
	All	30.65 (↑)	63.97 (↑)	64.87 (↑)
F1	3	0.40	0.62	0.64
	All	0.42 (↑)	0.63 (↑)	0.67 (↑)
Llama 3.1 8B				
R-L	3	17.67	12.30	15.03
	All	17.76 (↑)	13.75 (↑)	14.92 (↓)
BS	3	87.34	86.53	86.65
	All	87.54 (↑)	87.23 (↑)	87.06 (↑)
CS	3	31.82	25.16	30.72
	All	31.99 (↑)	27.47 (↑)	29.99 (↓)
C_{GT}	3	63.52	42.79	44.74
	All	64.47 (↑)	57.26 (↑)	62.04 (↑)
C_{LLM}	3	33.32	44.67	48.93
	All	27.75 (↓)	55.89 (↑)	58.86 (↑)
F1	3	0.39	0.43	0.46
	All	0.34 (↓)	0.56 (↑)	0.59 (↑)

Table 5: Ablation study with GPT 4o-mini + RAG and Llama 3.1 8B results in "coverage" (Coverage of Ground Truth Limitation (C_{GT}), LLM Generated Limitation (C_{LLM}), F1-score) and "performance" (**R**ouge-x, **B**ert**S**core (BS), and Cosine (CS) similarity in the **NeurIPS data**.

8 Conclusion

We present a new approach for automatically extracting, generating, and evaluating limitations in scientific articles. Our method explores incorporating cited works, accommodating top sections of the entire paper, and integrating review feedback to capture perspectives beyond those of the original authors. To evaluate the effectiveness of our system, we introduce a granular text evaluation framework that breaks down limitations into more minor points and employs LLMs as a Judge for assessing alignment. Human review validates our extraction and LLM-as-Judge pipeline, showing strong agreement with expert judgments.

Model	VD	C_GT	C_LLM	F1
GPT 4o-mini	Vanila k =3	40.34	49.34	0.43
GPT 4o-mini	LLM re-ranker	**68.84**	**64.87**	**0.67**

Table 6: Performance between different retriever approaches in VD (Vector Database) in RAG (vanilla RAG (considering top 3 chunks) vs LLM re-ranker)

Limitations

In this work, we focused on venues in natural language processing (ACL papers from 2023-2024) and machine learning (NeurIPS papers 2021-2022), and Biology domain papers from PeerJ, which ensures high relevance and quality but insufficient for broader generalizability. While this scope allows us to benchmark the performance of LLMs in extracting limitations from well-structured scientific texts, we acknowledge that the findings may not generalize to papers from other fields, such as social sciences, physics, chemistry, or mathematics where writing conventions and limitation styles may differ.

Due to high API costs, we did not experiment with GPT-4 or GPT-4o; instead, we opted for GPT-4o Mini as a cost-effective alternative. While we incorporated OpenReview comments for NeurIPS papers, we could not find them for ACL papers. Furthermore, we relied only on GPT-4o Mini as the evaluation judge and did not experiment with other LLMs for assessment. To evaluate the effectiveness of LLMs as both text extractors and judges, we conducted a human annotation study with 200 samples and only three annotators.

A key threat to validity is contamination bias, when evaluation examples (or close paraphrases) appear in a model's training data, artificially inflating performance. To guard against this, we tested whether GPT-4o mini had been trained on our NeurIPS 2021–2022 dataset by providing only each paper's title and prompting it to summarize the content and identify limitations. In every case, the model replied with a disclaimer indicating unfamiliarity with the specific work (e.g., "I am not familiar with the specific paper titled ..."). This consistent outcome suggests the model lacked prior exposure to the full texts, supporting the integrity of our evaluation.

While we selected GPT-4o mini for text extraction, generation, and evaluation due to its superior performance, relying on a single LLM for these roles introduces several potential biases. We took

specific steps to mitigate these risks: To counter self-validation bias, where the model might favor its own output, we cross-referenced its judgments with human evaluations and incorporated RAG. For positional bias, where the model may favor the first input when comparing texts, we swapped the input order to ensure consistent results. To reduce confirmation bias, the tendency to generate generic limitations, we used RAG to introduce more diverse evidence. Finally, to check for hallucinations, three human annotators verified that all extracted limitations were grounded in the source text. Although these strategies are crucial for improving reliability, we acknowledge that they do not completely eliminate these inherent biases.

For future work, we will expand our dataset to more diverse domains (e.g., bioinformatics, cognitive science) to test the cross-domain robustness of our models. We also plan to enhance our generation framework by exploring more advanced multi-agent and open-source LLMs via RAG. Finally, we will scale our human validation efforts with a larger, more diverse pool of expert annotators to enable a deeper and more reliable analysis.

Ethics Statement

This research adheres to ACL ethical standards. All data, including research papers and OpenReview feedback, were sourced from public repositories in compliance with their usage policies and were not filtered based on discriminatory attributes. Our user study involved three computer science graduate students who participated voluntarily with no conflicts of interest.

We acknowledge and address inherent LLM risks, including biases from training corpora, confirmation bias toward "safe" limitations, fluency and verbosity biases favoring longer or well-written outputs, and self-validation bias when using the same model for multiple tasks. To mitigate these, we (1) ground all generations in source content and peer reviews via a RAG framework to improve factuality and reduce verbosity; (2) diversify our ground truth by incorporating human-authored OpenReview critiques; (3) use a multi-model judge setup to break self-validation circularity; and (4) conduct parallel human evaluations to detect overconfidence and other model-specific biases. We recognize that further work is needed to rigorously quantify these issues and plan to investigate cross-domain robustness in future studies.

References

Ibrahim Al Azher. 2024. Generating suggestive limitations from research articles using llm and graph-based approach. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–3.

Ibrahim Al Azher and Hamed Alhoori. 2024. Mitigating visual limitations of research papers. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8614–8616. IEEE.

Ibrahim Al Azher, Venkata Devesh Reddy, Hamed Alhoori, and Akhil Pandey Akella. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *ACM/IEE Joint Conference on Digital Libraries (JCDL)*.

Thomas M Annesley. 2010. The discussion section: your closing argument. *Clinical chemistry*, 56(11):1671–1674.

Ibrahim Al Azher, Miftahul Jannat Mokarrama, Zhishuai Guo, Sagnik Ray Choudhury, and Hamed Alhoori. 2025. Futuregen: Llm-rag approach to generate the future work of scientific article. *arXiv preprint arXiv:2503.16561*.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Suzanne Bunniss and Diane R Kelly. 2010. Research paradigms in medical education research. *Medical education*, 44(4):358–366.

Lisa Chasan-Taber. 2014. *Writing dissertation and grant proposals: Epidemiology, preventive medicine and biostatistics*. CRC Press.

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

Jean Escande, Christophe Proust, and Jean Christophe Le Coze. 2016. Limitations of current risk assessment methods to foresee emerging risks: Towards a new methodology? *Journal of Loss Prevention in the Process Industries*, 43:730–735.

Kevin W Eva and Lorelei Lingard. 2008. What's next? a guiding question for educators engaged in educational research.

Abdur Rahman Bin Mohammed Faizullah, Ashok Urlana, and Rahul Mishra. 2024. Limgen: Probing the llms for generating suggestive limitations of research papers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 106–124. Springer.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Steven N Goodman, Jesse Berlin, Suzanne W Fletcher, and Robert H Fletcher. 1994. Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of internal medicine*, 121(1):11–21.

Sue Greener. 2018. Research limitations: the need for honesty and common sense.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Richard Horton. 2002. The hidden research paper. *Jama*, 287(21):2775–2778.

Ken Hyland. 1998. Hedging in scientific research articles.

John PA Ioannidis. 2007. Limitations are not properly acknowledged in the scientific literature. *Journal of clinical epidemiology*, 60(4):324–329.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *Transactions on Machine Learning Research*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

Victor M Montori, Roman Jaeschke, Holger J Schünemann, Mohit Bhandari, Jan L Brozek, PJ Devereaux, and Gordon H Guyatt. 2004. Users' guide to detecting misleading claims in clinical research reports. *Bmj*, 329(7474):1093–1096.

Kishore Papineni. 2001. Bleu: a method for automatic evaluation of mt. *Research Report, Computer Science RC22176 (W0109-022)*.

MA Puhan, N Heller, I Joleska, L Siebeling, P Muggensturm, M Umbehr, S Goodman, and G ter Riet. 2009. Acknowledging limitations in biomedical studies: The alibi study. In *The Sixth International Congress on Peer Review and Biomedical Publication*, pages 10–12. JAMA and BMJ Vancouver, Canada.

Milo A Puhan, Elie A Akl, Dianne Bryant, Feng Xie, Giovanni Apolone, and Gerben ter Riet. 2012. Discussing study limitations in reports of biomedical studies-the need for more transparency. *Health and quality of life outcomes*, 10:1–4.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Paula T Ross and Nikki L Bibler Zaidi. 2019. Limited by our limitations. *Perspectives on medical education*, 8:261–264.

Gerben Ter Riet, Paula Chesley, Alan G Gross, Lara Siebeling, Patrick Muggensturm, Nadine Heller, Martin Umbehr, Daniela Vollenweider, Tsung Yu, Elie A Akl, et al. 2013. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. *PloS one*, 8(11):e73623.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Jing Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 563–578.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Vilma Žydžiūnaitė. 2018. Implementing ethical principles in social research: Challenges, possibilities and limitations. *Profesinis rengimas: tyrimai ir realijos*, 1(29):19–43.

A Appendix

In our PointWise evaluation method, we measured precision, recall, and F1 score from True Positive, False Positive, and False Negative.

A.1 Coverage Measurement

We compute:

$$\mathrm{P}_{r_i} = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \quad \mathrm{R}_{r_i} = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i},$$

and the F_1 score is the harmonic mean of P_{r_i} and R_{r_i}.

A.2 Performance Measurement

A. Text-Based Evaluation. We apply standard text similarity metrics to each matched pair, including ROUGE-1, ROUGE-L, BERTScore, Cosine Similarity, Jaccard Similarity, and BLEU, calculating the number of overlapping unigrams, the longest sequence of words, and the similarity between contextual embeddings.

B. Keyword-Based Evaluation. We employ KeyBERT (Grootendorst, 2020) to extract a set of top keywords from the ground truth limitations K_{G_i} and from the LLM-generated limitations K_{H_i}. We then measure the cosine and Jaccard similarity between K_{G_i} and K_{H_i} for each paper P_i and average these scores across the dataset.

C. Heading-Based Evaluation. We also compare concise "headings" or short titles for each limitation. Let T_{G_i} be the heading for G_i and T_{H_i} the heading for H_i. We compute BERTScore between T_{G_i} and T_{H_i} for every paper P_i and then average these values. This provides a high-level measure of how closely the top-level concepts align.

By combining coverage and performance metrics in a PointWise manner, our framework provides a detailed assessment of how well an LLM-generated set of limitations captures the breadth and depth of the ground truth. This approach also facilitates fine-grained error analysis by examining matched pairs on a per-limitation basis.

We measure coverage for both ground truth and LLM-generated limitations *independently*, focusing on each unique limitation within the matched pairs.

Furthermore, we conduct experiments using:

1. The top three sections (*Abstract, Introduction, and Conclusion*)
2. The entire paper (*full paper*)

This setup enables us to examine how restricting the analysis to specific sections affects coverage and matching performance.

We used three distinct prompts to check the topic-level similarity between ground truth limitations and LLM-generated limitations (Figure 5, Appendix). To overcome the position bias, we choose the consistent one.

	GPT-4	Llama	HE1	HE2	HE3
GPT-4	1	0.71	0.9	0.92	0.95
Llama	-	1	0.81	0.79	0.76
HE1	-	-	1	0.98	0.95
HE2	-	-	-	1	0.97
HE3	-	-	-	-	1

Table 7: Evaluating how good LLM 'as a Judge' by checking Human Expert (HE) and model (GPT-4o mini, Llama-3.1 400B) agreement in determining whether an extracted limitation *matches* a generated one (in PointWise Evaluation).

Section	Cosine Similarity
Abstract vs Limitation	**33.27**
Introduction vs Limitation	**33.06**
Related Work vs Limitation	25.10
Methodology vs Limitation	26.58
Dataset vs Limitation	25.59
Conclusion vs Limitation	**33.04**
Experiment and Results vs Limitation	31.73

Table 8: Cosine Similarity between each section and the Limitation section.

Prompt = '''
Here is the text containing extracted limitations. Please identify and list each limitation, ensuring that each one addresses a distinct topic or point. '''

Figure 3: Prompt to extract limitations from ground truth text.

Prompt = '''
You are a helpful, respectful, and honest assistant for generating limitations or shortcomings of a research paper. I am providing 'Abstract', 'Introduction', 'Related Work', 'Methodology', 'Experiment and Results', 'Conclusion', and other sections of a scientific paper alongside the related cited papers texts. Generate limitations based on these texts. '''

Figure 4: Prompt to generate limitations from Input and cited papers text.

Prompt 1 = '''
A tuple contains (list1, list2). Check whether both 'list1' and 'list2' have similar topics or limitation. If both have similar topics or limitations you can say "Yes", otherwise "No". Your answer should be "Yes" or "No" with explanation. '''

Prompt 2 = '''
A tuple contains (list2, list1). Check whether both 'list2' and 'list1' have similar topics or limitation. If both have similar topics or limitations you can say "Yes", otherwise "No". Your answer should be "Yes" or "No" with explanation.

Prompt 3 = '''
Check whether 'list2' contains a topic or limitation from 'list1' or 'list1' contains a topic or limitation from 'list2'. Your answer should be "Yes" or "No" with explanation.

Figure 5: LLM as a Judge for each limitation. We use three distinct prompts to verify consistency.

Ground Truth	Gen. Lim.	GPT 4	Llama 3	User
Model Complexity Concerns: - There is a question regarding whether the performance gains are due to increased model complexity rather than the proposed recursive mixing approach.	**Potential for Increased Complexity**: Although the method is described as simple, the recursive nature of the approach may introduce complexity in implementation and understanding, particularly for practitioners who may not be familiar with the underlying concepts	Yes	Yes	Yes
Computational Complexity: The Dual-aspect Attention mechanism significantly increases computational complexity, which may affect the performance and practicality of the proposed method.	'**Complexity of the Model**: The DACT arch, while innovative, introduces additional complexity compared to traditional models', 'This complexity may lead to longer training times and increased computational resource requirements, which could be a barrier for practical applications in resource-constrained environments'	Yes	Yes	Yes
Insufficient Detailed Experimental Analysis: The paper primarily provides quantitative results without a detailed experimental analysis, which limits the understanding of the findings.	**Limited Discussion on Failure Cases**: While the paper discusses success and failure cases of existing methods, it may not provide sufficient insight into the specific conditions under which SCILL might fail, limiting the understanding of its robustness.	Yes	No	Yes
Connection to Practical Methods: There is an unknown connection between the theoretical findings and popular ensemble methods used in practice, which raises questions about the practical relevance of the work.	**Unexplored Variants of Ensemble Methods**: The research does not explore other ensemble methods beyond those mentioned, such as boosting or stacking, which may have different theoretical properties and could provide additional insights into the interpolating regime	No	Yes	Yes
Lack of Exploration of Alternative Methods: The paper does not sufficiently explore or compare the performance of other methods like MFVI and SGLD in the context of covariate shift, which could provide a more comprehensive understanding of the problem.	**Absence of Dataset Details**: The dataset section is marked as "nan," indicating a lack of information about the datasets used for experimentation', 'This omission makes it difficult to assess the validity and applicability of the findings, as the choice of dataset can significantly influence the results	No	No	No

Table 9: Examples of Annotator, GPT 4o-mini, and LLama judgement on whether a generated limitation should be matched with a ground-truth limitation or not.

Ground Truth Limitations: Our goal is to provide comprehensive study of learning algorithms, including their hyperparameters. However, our study is limited on the variety of other design choices to restrict the experimental complexity. While we studied both synthetic and realistic image datasets, both these datasets are relatively simple with the same small number of generative factors and each of the factor follows uniform distribution. For learning algorithms, we focus on studying the inductive bias on the representation format while fixing the model arch design which can impact the results. Moreover, we did not study hyperparameters beyond those related to the latent representations. Specifically, we did not study how the type and configurations of the optimizer and the batch size would change the results; instead, we followed common setups in previous studies.
' ' '

LLM Extracted Limitations:
1. Here are the limitations extracted from the provided limitation:,
2. **Experimental Complexity**: The study is limited in the variety of design choices to reduce experimental complexity.,
3. **Dataset Complexity**: The study utilizes relatively simple synthetic and realistic image datasets, which have a small number of generative factors.,
4. **Generative Factors**: Each generative factor in the datasets follows a uniform distribution, limiting the diversity of the data.,
5. **Model arch**: The focus is on studying the inductive bias related to the representation format while keeping the model arch design fixed, which may influence the results.,
6. **Hyperparameter Scope**: The study does not explore hyperparameters beyond those related to latent representations.,
7. **Optimizer and Batch Size**: The impact of different types and configurations of optimizers and batch sizes on the results is not examined, as the study adheres to common setups from previous research.

Figure 6: Ground Truth Limitations and LLM Extracted Limitations in NeurIPS dataset.

Ground Truth Limitations: While our work provides useful starting point for understanding student feedback,there are limitations to our work. Addressing these limitations will be an important area for future research. Comments may not reflect real student feedback. The comments in our dataset are from users who have chosen to post publicly on youtube. Addi- tionally,the comments may include features specific to this online education setting. Thus,the comments may reflect real student comments from these courses. There is selection bias in lecture sources. Sight includes lectures that may be drawn from the most successful offerings of that course. The instructional quality may not be representative of typical instruction. Thus,inferences drawn about the instruction should be interpreted with caution,as they might not generalize to other lecture settings. We analyze only english comments. We analyze only english comments because the lecture content is given in english and the authors are most comfortable with english. As result,our rubric may not capture the types of feedback from nonenglish students watching lectures taught in english. we annotate small subsample of the data to assess the validity of the automatic labels,we conduct diagnostic study on small,randomly selected subset of the dataset,comprising approximately of the comments. Our work aims to establish preliminary evaluation of the humanmodel agreement and model annotations,and further validation of the automatic labels is necessary.
' ' '

LLM Extracted Limitations: Here are the limitations extracted from the provided limitation:

1. **Comments May Not Reflect Real Student Feedback**: The dataset consists of comments from users who have chosen to post publicly on YouTube, which may not accurately represent genuine student feedback.
2. **Features Specific to Online Education Setting**: The comments may include characteristics that are unique to the online education context, potentially skewing the interpretation of student feedback.
3. **Selection Bias in Lecture Sources**: The lectures included in the analysis may be drawn from the most successful offerings of a course, leading to a selection bias that does not represent typical instructional quality.
4. **Generalizability of Instructional Quality**: Inferences about instructional quality based on the selected lectures should be made with caution, as they may not generalize to other lecture settings.
5. **Analysis Limited to English Comments**: The analysis is restricted to comments in English, which may overlook feedback from non-English speaking students who are watching lectures taught in English.
6. **Small Subsample for Validity Assessment**: The study annotates a small subsample of the data to assess the validity of automatic labels, which may not provide a comprehensive evaluation of the entire dataset.
7. **Need for Further Validation of Automatic Labels**: The work establishes a preliminary evaluation of human-model agreement and model annotations, indicating that further validation of the automatic labels is necessary.

Figure 7: Ground Truth Limitations and LLM Extracted Limitations in ACL dataset.

OpenReview:
Reviewer 1: This paper tries to tackle the domain generalization problem by finding flat minima. It is interesting to investigate the loss landscape of the DNNs on domain generalization problems. Originality: The method proposed in this paper is mainly built on previous approaches such as SWA, and the authors mention this in the submission. Quality: This submission is complete piece of work and the claims are well supported by the empirical results. The evaluation is reasonable. Clarity: This submission is well-written and easy to follow. Significance: This paper provides interesting empirical observations and improves the state-of-the-art by applying their proposed new approach. Pros: The proposed method consistently improves the model performance on domain generalization problems, and the proposed method is flexible and can be easily integrated with other training approaches for better performance. Cons: The generalization bound does not provide intuition for robust risk minimization..

Reviewer 2: Originality The paper is, perhaps by design, not particularly original. Rather, it makes the observation that an existing set of techniques with particular goal also has applications to different goal. This is still, to the best of my knowledge, novel observation and contribution, and furthermore the modifications made to SWA to arrive at SWAD also appear novel. Nevertheless, these modifications are relatively small, so the method can still be viewed as version of the general idea of SWA, and thus originality is not the paper' strong suit. Quality The paper is of relatively high quality, primarily due to the empirical results. did not carefully check the theory for correctness, but am unsure as to the significance of the theorems. They appear rather similar to prior theoretical results on how flat minima can lead to better generalization, but with some additional divergence terms thrown in as we are dealing with the DG setting...

Reviewer 3: Pros: The paper tackles an important and relevant problem to the ML community. The proposed algorithm is relatively simple to implement in practice and perform well experimentally. Cons: My major concern is wrt to the novelty of the work and the connection/motivation between the proposed algorithm SWAD, domain generalization and the derived theoretical insights...
' ' '

LLM extracted OpenReview:
"1. . **Connection Between Theory and Method**: The relationship between the theoretical framework presented in the paper and the SWAD algorithm is weak. There is a need for a clearer connection between the flatness term in the equations and the algorithm's performance.",
'2. . **Novelty of Theoretical Results**: The novelty of the theorem compared to classic results, such as those from Ben-David, is questioned. The theoretical contributions do not significantly advance the understanding of flat minima in the context of domain generalization.',
'3. . **Generalization Bound Intuition**: The generalization bound provided does not offer sufficient intuition for robust risk minimization (RRM), and it does not convincingly demonstrate that RRM is superior to standard empirical risk minimization (ERM).'],
'4. . **Dependence on Previous Work**: The proposed method heavily relies on the existing SWA approach, with only minor modifications. This raises concerns about the originality of the contribution.',
'5. . **Theoretical Framework Limitations**: The reliance on VC dimension in the theoretical framework is problematic, as it can be excessively large for the models used, especially in domain generalization scenarios.',
'6. . **Overlooked Details in Proofs**: There are concerns that important terms have been overlooked in the technical proofs, which could undermine the validity of the results.',
'7. . **Lack of Clear Motivation for SWAD**: The motivation for using SWAD over SWA is not clearly articulated, making it difficult to understand the advantages of the proposed algorithm.',
'8. . **Introduction of New Hyperparameters**: The proposed algorithm introduces additional hyperparameters, which may complicate the implementation and tuning process.',
'9. . **Limited Societal Impact Discussion**: While a limitations section is included, there is a lack of discussion regarding the societal impacts of the research, although specific concerns were not identified.',
"10. . **Potential for Broader Applications**: The method's applicability beyond domain generalization is not explored, which could limit its significance and impact in the broader machine learning community.

Figure 8: Tool extracted OpenReview and LLM Extracted OpenReview.