Exploring Evidence-Grounded Review Generation with Multimodal Inline RAG

Anonymous ACL submission

Abstract

Multimodal review generation is essential in academic research. However, existing LLM-based models often struggle to pinpoint specific details within papers, yielding generalized, unsupported critiques that compromise final decisions. To address these limitations, we propose **AnchorReviewer**, a coarse-to-fine, evidence-grounded multimodal review generation framework. Central to our architecture is an integrated pipeline featuring an Inline Retrieval-Augmented Generation (InRAG) module. Combined with our Unified Embedding Module, InRAG facilitates the high-precision extraction of multimodal evidence chunks, ensuring our detailed reviews are strictly anchored to the original content. Furthermore, we incorporate a Reflection Mechanism to refine initial outputs and maximize decision accuracy. To support model evaluation and our Progressive Training Strategy, we construct the large-scale **AnchorReview** dataset, comprising diverse multimodal review samples. Experimental results demonstrate that AnchorReviewer significantly improves review reliability, excelling in critical evaluation, multimodal understanding, and decision consistency. Our code and dataset will be publicly available.

1 Introduction

Leveraging exceptional capabilities in language understanding and generation, large language models (LLMs) increasingly drive innovations in academic peer review, demonstrating widely recognized potential (Naddaf, 2025; Thakkar et al., 2025). By efficiently processing, summarizing, and reasoning over massive corpora, LLMs effectively assist researchers in initial manuscript evaluation and filtering. This capability significantly alleviates individual human judgment burdens while maximizing research efficiency and the reliability of academic dissemination (Stelmakh et al., 2021; Fox et al., 2023; Liang et al., 2024).

Current efforts to advance LLM-based review systems primarily explore two paradigms: agent-based orchestration (Lu et al., 2025; Jin et al., 2024; Zhang et al., 2026; Lu et al., 2024)and data-driven fine-tuning (Zhu et al., 2025b; Weng et al., 2025). While agent-based frameworks simulate human workflows, their reliance on commercial APIs risks manuscript data leakage (Zhu et al., 2025a), incurs massive token costs (Baumann et al., 2026), and fatally yields inflated scores alongside generic critiques (Rao et al., 2025). Consequently, researchers have turned to data-driven fine-tuning, training dedicated reviewer models on domain-specific corpora to circumvent privacy hazards through local deployment. However, constrained by their text-only architectures, these fine-tuned reviewers produce homogenized reviews that fail to articulate targeted insights into substantive innovations or defects.

We argue that the bottleneck of data-driven fine-tuning fundamentally stems from a "visual evidence vacuum" driven by a complete absence of multimodal perception. Because existing systems cannot precisely anchor critiques to dense visual layouts—such as complex equations or rigorous experimental charts—they lack the capacity to execute substantive verification of the underlying data. Lacking concrete visual evidence to refute textual claims, agents passively conform to the authors' narratives, forging a superficial "compromising consensus" that directly inflates evaluation scores. Simultaneously, purely text-based fine-tuned models remain entirely blind to the core visual evidence defining academic contributions, completely severing the logical link between the review text and specific experimental results. Ultimately, this modality blind spot prevents current systems from delivering the rigorous, evidence-grounded micro-scrutiny characteristic of human experts.

To overcome these issues, we propose AnchorReviewer, a lightweight, end-to-end multimodal reviewing architecture. Mimicking the human

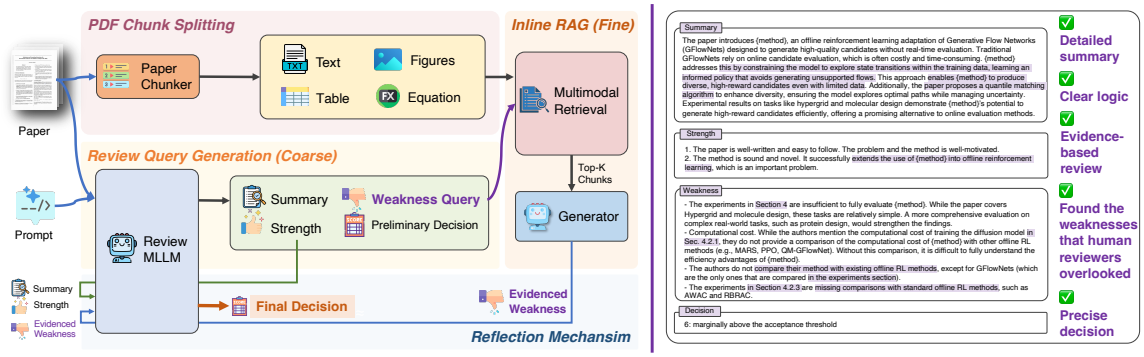

Figure 1: **Left**: We propose AnchorReviewer, a coarse-to-fine, evidence-grounded review pipeline: an MLLM generates weakness queries, RAG retrieves relevant multimodal evidence, and the model produces detailed reviews. A reflection step refines final decisions for accuracy and reliability. **Right**: In real review cases, it generates detailed, evidence-grounded reviews with strong logical consistency and accurate scores.

"glance-then-scrutinize" cognitive process, we decouple the pipeline into Coarse, InRAG, and Reflection stages to entirely prevent brute-force document appending. Our Coarse stage extracts macro-level critiques from global contexts while simultaneously emitting precise local queries for micro-level defects. These queries trigger our Inline RAG (InRAG) module, leveraging a unified embedding space (UniEmb) to tightly anchor high-fidelity multimodal evidence chunks (e.g., figures, equations). Ultimately, our Reflection Mechanism synthesizes these clues within a noise-free context to yield rigorous, evidence-grounded detailed reviews.

We underpin our framework's critical depth by constructing AnchorReview, the first large-scale multimodal reviewing corpus explicitly annotated with paper-region (bounding-box) evidence locations. Curated from top-tier AI conferences (ICLR, NeurIPS, and ICML), AnchorReview is hierarchically organized into three nested subsets that align with our coarse-to-fine paradigm: a General subset of 68.4K papers / 300.7K paper–review pairs, an Evidenced pool of 31.2K papers segmented into 940.3K multimodal chunks (text, figures, tables, equations), and an Anchored subset of 90K fine-grained query–evidence–weakness triples in which every critique is anchored to a specific bounding box on the original PDF. Leveraging this hierarchically annotated data for progressive fine-tuning, our 6B model is immunized against the positivity bias, enforcing a strict mapping between stringent textual critiques and concrete visual evidence.

In summary, our contributions are threefold.

- We propose **AnchorReviewer**, an evidence-grounded multimodal review framework that unifies a Coarse-to-Fine pipeline, InRAG (Inline Retrieval-Augmented Generation) for query-triggered fine-grained retrieval over a UniEmb unified multimodal embedding space, and a Reflection mechanism for evidence-consistent decision refinement.

- We construct **AnchorReview**, the first large-scale peer-review corpus that jointly provides multi-modality coverage and paper-region (BBox) grounding, hierarchically organized into three nested subsets that directly drive our progressive training.

- Experimental results demonstrate that the proposed model reliably identifies evidence and generates accurate decisions, achieving SOTA performance across benchmarks.

2 Related Works

2.1 LLM-based Review Generation

Automated review generation leverages LLMs through structured reasoning (Zhu et al., 2025b; Wu et al., 2025), hierarchical decomposition for complex arguments (Chang et al., 2025), and multi-agent orchestration for collaborative verification (D'Arcy et al., 2024; Taechoyotin and Acuna, 2025; Gao et al., 2024). Concurrently, extensive benchmarks evaluate these evolving capabilities (Li et al., 2025; Shin et al., 2025; Thakkar et al., 2025). However, a critical limitation persists: these predominantly text-centric models fundamentally overlook essential visual evidence. Consequently, this modality gap produces generic feedback, lacking the fine-grained, evidence-grounded insights characteristic of expert human reviews.

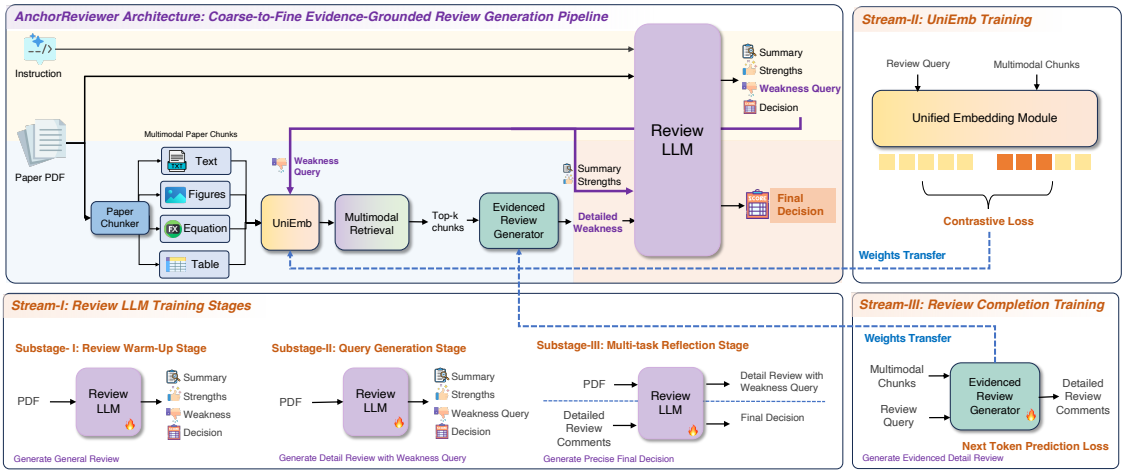

Figure 2: Left-top: The Coarse-to-Fine Evidence-Grounded Pipeline improves quality by linking reviews to evidence. It splits papers into multimodal chunks, generates queries, retrieves evidence via RAG, and drafts detailed reviews. A reflection mechanism ensures logical consistency.Others: The Parallel Progressive Training Strategy utilizes three concurrent streams and three progressive stages to efficiently train the Review LLM.

2.2 Multimodal Retrieval-Augmented Generation

To overcome unimodal limitations, Multimodal Retrieval-Augmented Generation (RAG) integrates cross-modal knowledge through unified reasoning architectures like UniversalRAG (Yeo et al., 2025) and VisRAG (Yu et al., 2025). To refine visual-textual alignment, mechanisms such as mR2AG (Zhang et al., 2024) employ reflective feedback loops, while MLLM-driven rerankers (Liu et al., 2025; Chen et al., 2025; Yue et al., 2025) significantly enhance retrieval relevance. Additionally, specialized frameworks like BayesRAG (Li et al., 2026) and Video-RAG (Luo et al., 2024) utilize probabilistic corroboration and visual alignment to resolve complex semantic isolation.

Despite these advancements, existing Multimodal RAG systems struggle with localized evidence grounding within dense professional documents. The inability to precisely cite minute visual-textual details frequently yields generated reviews that lack the rigorous, incisive evidence chains essential for high-stakes academic evaluation.

3 AnchorReviewer

Overview Figure 2 illustrates the overall framework of AnchorReviewer: the top-left portion depicts the Coarse-to-Fine Evidence-Grounded Review Generation Pipeline, and the remaining portion summarises the Parallel Progressive Training Strategy. The paper PDF is segmented into multimodal evidence chunks (text, figures, tables), and the LLM generates an initial weakness query from the original paper. Using RAG, the model retrieves relevant evidence, combining it with the query to generate a detailed review. Finally, the reflection mechanism refines the review by integrating the weaknesses and initial feedback, producing a final, evidence-supported, and consistent decision.

3.1 Coarse-to-Fine Evidence-Grounded Review Generation Pipeline

AnchorReviewer employs a Coarse-to-Fine Evidence-Grounded Review Generation Pipeline, optimizing review generation by aligning it with literature evidence to improve accuracy and verifiability. We first represent the paper as a multimodal chunk set:

$$\mathcal{D} = \{c_1, c_2, \ldots, c_N\}, \tag{1}$$

where each c_i is an evidence unit produced by the chunker—a text paragraph, figure, table, or display equation—so that different modalities carry distinct information. A pre-trained large language model M_{query} then reads the entire paper and emits a set of preliminary weakness queries $Q = \{q_1, \ldots, q_K\} = M_{\text{query}}(\mathcal{D})$, which localise potential weak points and serve as anchors for subsequent retrieval.

For every query q_k, a Retrieval-Augmented Generation (RAG) mechanism returns the subset of evidence chunks most relevant to it, $E(q_k) = \text{RAG}(q_k, \mathcal{D})$, so that the generated content cites authentic excerpts from the paper rather than relying on pre-trained priors. Merging the per-query results gives the unified evidence set $E = \bigcup_{k=1}^{K} E(q_k)$,

which is then fed, together with Q, into the generator G_{weakness} to produce a detailed weakness assessment $W = G_{\text{weakness}}(Q, E)$.

Finally, the decision module M_{decision} consumes the detailed weaknesses and their evidence and emits the final review decision $D = M_{\text{decision}}(W, E)$, so that the decision can be traced back to its supporting evidence. By chaining query generation, evidence retrieval, and fine-grained writing into a single pipeline, every step of the review is grounded on structured evidence from the paper itself.

3.2 Inline Retrieval Augment Generation (InRAG)

We propose Inline Retrieval-Augmented Generation (InRAG) to support precise evidence retrieval and generation inside the Coarse-to-Fine Pipeline. By embedding all modalities into a single semantic space, InRAG enables consistent cross-modal retrieval and grounded generation.

Unified Embedding for Multimodal Chunks The chunks $\mathcal{D}$ from Eq. 1 span text, figures, tables, and other modalities, and important information frequently resides in figures or tables. Because the embedding gap across modalities degrades both retrieval and grounded generation, we introduce a unified embedding model E_{uni} that maps every modality into a shared semantic space,

$$z_i = E_{\text{uni}}(c_i), \quad i = 1, \ldots, N. \tag{2}$$

The same encoder is applied to natural-language queries, so similarities between any query and any chunk can be measured directly in this space, closing the cross-modal feature gap and enabling effective retrieval across text, figures, tables, and equations.

Multimodal Evidence Retrieval For each weakness query $q_k \in Q$, we compute the similarity between the query embedding and every chunk embedding in the unified space, keeping the top-k most relevant chunks:

$$E_{\text{inline}}(q_k) = \text{Top}_k\Big(\text{sim}(E_{\text{uni}}(q_k), E_{\text{uni}}(c_i)) \Big| c_i \in \mathcal{D}\Big), \tag{3}$$

where $\text{sim}(\cdot, \cdot)$ is the similarity metric (e.g. cosine similarity) in the unified space, and the resulting Top-k chunks serve as the retrieval context for that query.

Evidence-Grounded Review Generation We feed each query together with its retrieved evidence context into the weakness generator to obtain the detailed weakness $w_k = G_{\text{weakness}}\big(\{q_k\} \cup E_{\text{inline}}(q_k)\big)$; aggregating across all queries gives the InRAG output $W_{\text{inline}} = \{w_k\}_{k=1}^{K}$. Treating the retrieved chunks as additional in-context evidence forces the generated review to rely on both pretrained knowledge and document-level evidence, which sharpens both the precision and the semantic depth of the identified weaknesses.

Unlike traditional RAG, which decouples retrieval from generation, InRAG interleaves the two on a per-query basis. This yields higher-quality, traceable detailed weaknesses for the subsequent reflection and decision stages, and scales more gracefully to long and complex documents.

3.3 Reflection Mechanism for Review Refinement

To improve review accuracy and consistency, we propose a Reflection mechanism that re-inputs the summary and strengths into the LLM after generating the detailed weakness, refining the final review decision and comments.

We input the summary, strengths, and generated weaknesses $\{w_1, w_2, \ldots, w_K\}$ into the previously used LLM M_{decision} to generate the final review decision.

$$D = M_{\text{decision}}(W, \text{summary}, \text{strength}). \tag{4}$$

Here, M_{decision} is the decision-generation model, which takes as input the previously generated detailed weakness W, along with the model's output summary and strengths, and produces the refined decision D.

3.4 Parallel Progressive Training Strategy

AnchorReviewer is optimized with three concurrent streams aligned with the three nested subsets of AnchorReview, as depicted on the remaining portion of Figure 2 (outside the top-left pipeline block).

Stream-I trains the reviewer with a coarse-to-fine curriculum (warm-up → query generation → multi-task reflection) on AnchorReview-General and -Anchored. Stream-II trains UniEmb on (query, chunk) pairs from AnchorReview-Evidenced with in-paper hard negatives. Stream-III closes the retrieve-and-generate loop on AnchorReview-Anchored. All language objectives use next-token

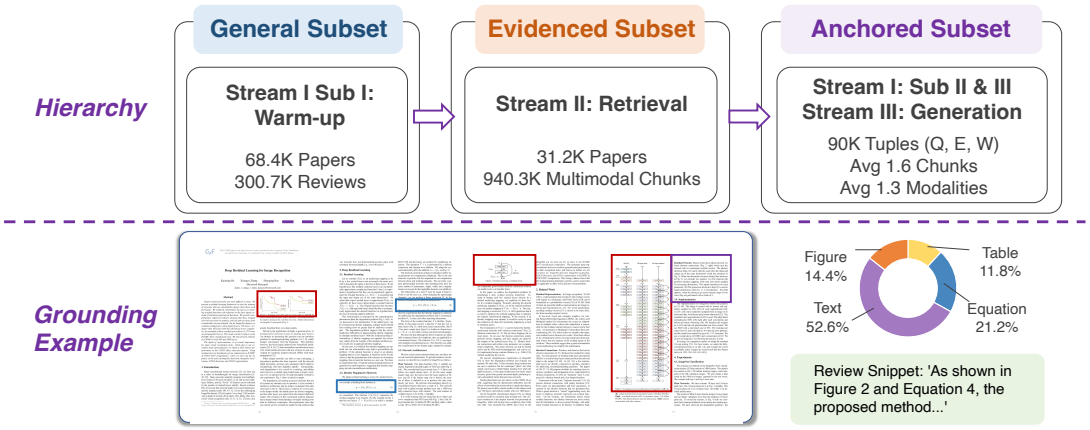

Figure 3: Dataset hierarchy and multimodal grounding example. Top: The progressive data construction pipeline and corresponding subset statistics. Bottom: A fine-grained grounding example illustrating modality distributions and the precise alignment between review snippets and multimodal document elements.

prediction; the explicit losses, per-stream data sources, and the progressive schedule on the loss weights $(\lambda_1, \lambda_2, \lambda_3)$ are deferred to Appendix C.

4 AnchorReview Dataset

To support our coarse-to-fine multimodal review pipeline, we construct AnchorReview, a large-scale peer-review corpus distinguished by two properties absent from prior resources:(1)multimodal evidence chunking (text, figures, tables, and equations) and (2)layout-level grounding that anchors every fine-grained weakness to a bounding box on the original PDF. The corpus is organized as three hierarchically nested subsets—General, Evidenced, and Anchored—which directly serve the three training streams of AnchorReviewer. To our knowledge, AnchorReview is the first peer-review corpus that jointly provides multimodal evidence chunking, layout-level grounding, and hierarchical organization, as summarized in Table 1.

4.1 Construction Pipeline

Figure 3 illustrates the construction pipeline, which proceeds in three stages and yields three increasingly fine-grained subsets.

Source and Chunking. We crawl all publicly accessible OpenReview submissions from ICLR, NeurIPS, and ICML, together with their reviewer turns. After de-duplicating multi-version revisions, discarding withdrawn or desk-rejected submissions, and removing reviews shorter than 50 tokens, we retain 68.4K papers and 300.7K reviews, forming the GENERAL subset. For papers whose reviews contain at least one locatable evidence reference (e.g., explicit section, figure, table, or equation citations), we further apply PP-DocLayout-V2 to detect the bounding boxes of sections, figures, tables, and equations; text regions are extracted via PaddleOCR, figure/table captions are matched by pattern-based heuristics, and equations are preserved as cropped images. This produces 940.3K multimodal chunks from 31.2K papers, constituting the Evidenced pool. Each chunk is represented as $c_i = (\text{type}_i, \text{bbox}_i, \text{content}_i, \text{caption}_i)$ and stored alongside a JSON configuration file.

Hierarchical Subsets. The three subsets supply complementary supervision signals to the parallel training streams of AnchorReviewer.

- AnchorReview-General (68.4K papers / 300.7K reviews) provides broad paper–review pairs and supervises the review warm-up substage of Substage-I, teaching the model the general syntax and rhetoric of peer review.
- AnchorReview-Evidenced (31.2K papers / 940.3K multimodal chunks) serves as the retrieval corpus for the Unified Embedding training (Stream-II) and as the in-context evidence pool at inference time.
- AnchorReview-Anchored (90K query–evidence–weakness triples), in which each triple (q, e, w) pairs a weakness query q, its grounding evidence e (one or more bounding-box chunks), and the corresponding detailed weakness w, is the most fine-grained subset. It directly drives the review-completion training (Stream-III) and additionally supervises the query-generation and multi-task reflection sub-stages of Substage-II and Substage-III.

Dataset	#Papers	#Reviews	Modalities	Review Sent. Labels	Paper Anchoring
PeerRead(Kang et al., 2018)	14.7K	10.7K	Text	×	×
ASAP-Review(Yuan et al., 2022)	28.0K	28.0K	Text	✓	×
DISAPERE(Kennard et al., 2022)	3.8K	5.0K	Text	✓	×
NLPEER(Dycke et al., 2023)	5.0K	11.0K	Text	×	partial
ReviewMT(Tan et al., 2024)	26.8K	92.0K	Text	×	×
AgentReview(Jin et al., 2024)	0.5K	10.5K	Text	×	×
DeepReview-13K(Zhu et al., 2025b)	13.4K	13.4K	Text	×	×
Re^2(Zhang et al., 2025)	19.9K	70.6K	Text	×	×
AnchorReview (Ours)	**68.4K**	**300.7K**	**Text + Fig + Tab + Eq**	✓	✓ **(BBox)**

Table 1: Comparative summary of peer-review datasets. AnchorReview is the only corpus that jointly provides multi-modality coverage and paper-region (bounding-box) anchoring.

Evidence Annotation. To build the ANCHORED subset, we decompose each detailed review into atomic sub-bullets and prompt GPT-5.1 with a fixed template (Appendix H) to extract, for every sub-bullet, a structured *(query, location reference)* pair. Location references such as "Table 3", "Section 4.2", or "Figure 2 " are then matched to chunks through regular expressions over chunk captions, numbers, and section headers; multi-chunk references are concatenated and treated as a single composite evidence. Sub-bullets whose location is generic (e.g., "the experiments") or whose match score falls below a confidence threshold are discarded, ensuring that every retained triple is anchored to a specific spatial region of the source PDF.

4.2 Quality Control

We apply a three-level quality-control pipeline to suppress noise: (i) automatic filtering of low-quality and suspected machine-generated reviews; (ii) cross-model validation, in which an independent judge (Claude Opus 4) re-annotates a 5% random sample of Anchored triples, yielding 98.6% cross-model agreement; and (iii) human verification, in which three NLP graduate students double-annotate 300 randomly sampled triples on query validity, evidence relevance, and spatial accuracy, yielding 92.4% overall validity with Cohen's $\kappa = 0.81$. These results indicate that the GPT-5.1-assisted anchoring is reliable at scale. Full procedures, scope, and per-stage statistics are provided in Appendix A.1; the human-evaluation guideline is in Appendix I.

4.3 Statistics and Analysis

We highlight two key observations from our corpus. First, the EVIDENCED pool averages 30.1 chunks per paper (52.6% Text, 21.2% Equation, 14.4% Figure, 11.8% Table). This near-even split between text and non-text evidence underscores the absolute necessity of multimodal grounding. Second, each ANCHORED weakness grounds on an average of 1.6 chunks across 1.3 modalities, proving that high-quality critiques routinely cross visual-textual boundaries entirely missed by prior text-only datasets. Appendix A details comprehensive dataset statistics.

5 Experiments

5.1 Experiment Setup

Implementation details. We use Qwen3-VL-2B(Bai et al., 2025) as the MLLM and generator, and Qwen3-VL-Embedding as the Unified Embedding model.Training follows our three-stream strategy for one epoch per stream with Adam.The learning rates for Stream-I are 1e-4/5e-5/5e-5 across its three substages; Stream-II/III use 1e-4 with cosine decayThe batch size is 16, all modules are fully fine-tuned, and each PDF page is rendered at 1081×1400. We use the ICLR portion of AnchorReview as the training set.

Evaluation Protocol. For in-domain evaluation, we use a paper-level held-out split of 300 ICLR'24/'25 paper–review pairs that do not appear in training. For out-of-distribution evaluation (§5.4), we additionally evaluate on NeurIPS and ICML papers, which appear in no fine-tuned reviewer's training data. Ground-truth ratings and decisions are taken from the publicly released OpenReview meta-reviews, ensuring evaluation labels are independent of any model-generated annotation in AnchorReview.

Baselines and Metrics. We compare two categories of baselines: (1) agent-style prompt baselines, including AI Scientist and AgentReview with

Method	Model	Decision Metrics		Rating Metrics			
		Accuracy ↑	F1 ↑	MAE ↓	MSE ↓	Match ↑	±1 Acc ↑
AgentReview(Jin et al., 2024)	Claude-Sonnet-4	0.6328	0.4786	1.7124	4.2861	0.1312	0.5164
	Gemini-2.5-Flash	0.6695	0.5021	1.5847	3.7649	0.1679	0.5458
AI Scientist(Lu et al., 2024)	DeepSeek-V3.2	0.7053	0.5556	1.6632	3.9789	0.0895	0.5316
	GPT-4o-mini	0.6513	0.4603	1.6051	3.9128	0.1436	0.5385
	GPT-5.2	0.6131	0.3984	1.9795	5.5385	0.0821	0.4205
CycleReviewer(Weng et al., 2025)	8B	0.6103	0.3968	1.7958	5.5445	0.2880	0.4241
DeepReviewer(Zhu et al., 2025b)	14B	0.5897	0.3651	1.4536	4.1031	0.3660	0.5309
AnchorReviewer (ours)	6B	**0.7744**	**0.6508**	**1.1641**	**2.6308**	**0.4051**	**0.5744**

Table 2: Main comparison of quantitative review metrics.

Method	Model	Weaknesses Score (0-5)					Review Score (0-5)					
		Overall	Consis.	Accuracy	Complete.	Diversity	Overall	Consis.	Accuracy	Complete.	Non-Contra.	Dec-Cons.
AgentReview	Claude-Sonnet-4	2.69	2.18	3.13	2.45	2.96	3.42	3.20	3.72	3.05	4.55	2.39
	Gemini-2.5-Flash	2.50	2.00	2.90	2.30	2.74	3.30	3.10	3.63	2.92	4.51	2.45
AI Scientist	DeepSeek-V3.2	2.72	2.20	3.14	2.58	2.97	3.43	3.22	3.74	3.10	4.57	2.58
	GPT-4o-mini	2.67	2.31	3.14	2.27	2.98	3.39	3.27	3.84	2.98	4.60	2.27
	GPT-5.2	2.66	2.16	3.10	2.42	2.96	3.40	3.18	3.71	3.02	4.54	2.41
CycleReviewer	8B	2.10	1.75	2.54	1.93	2.16	3.15	2.86	3.43	2.59	4.52	2.37
DeepReviewer	14B	2.43	1.78	2.71	2.57	2.67	3.38	3.13	3.68	3.23	4.55	2.31
AnchorReviewer	6B	**2.80**	**2.45**	**3.18**	**2.92**	**3.14**	**3.49**	**3.29**	**3.87**	**3.33**	**4.63**	**2.74**

Table 3: Main comparison of the review using GPT scoring.

DeepSeek-V3.2, GPT-4o-mini, GPT-5.2, Claude-Sonnet-4, and Gemini-2.5-Flash; (2) fine-tuned reviewer models, including CycleReviewer-8B and DeepReviewer-14B. We report decision metrics (Accuracy, F1), rating metrics (MAE, MSE, exact-match accuracy, and ± 1 accuracy), and GPT-judge review-quality metrics. Detailed metric definitions are provided in Appendix B.

5.2 Main Results

Decision and rating performance. As shown in Table 2, AnchorReviewer with only 6B parameters achieves the best score on every decision and rating metric, surpassing the strongest agent baseline by 6.91 Accuracy points and the larger DeepReviewer-14B by 18.47 Accuracy points. The simultaneous gains on Accuracy/F1 and MAE/MSE indicate that the coarse-to-fine multimodal pipeline improves decision reliability and score calibration jointly, rather than trading one for the other.

Review-quality evaluation. Table 3 shows that AnchorReviewer obtains the highest GPT-judge score on all 11 weakness and review dimensions. The largest gains are on weakness Completeness/Diversity and review Decision-Consistency, reflecting that grounding reviews to retrieved evidence yields more specific, calibrated, and actionable feedback.

Variant	Decision		Score			
	Acc. ↑	F1 ↑	MAE ↓	MSE ↓	Match ↑	±1 Acc. ↑
Qwen3-VL-2B (zero-shot)	0.4821	0.3245	2.1842	6.0157	0.0612	0.4137
+ Substage-I (warm-up)	0.6821	0.5079	1.3158	3.3474	0.3789	0.5526
+ Substage-II (query gen.)	0.7283	0.5794	1.2399	2.9891	0.3920	0.5635
+ InRAG & general reflection	0.7388	0.5512	1.2051	2.7512	0.3845	0.5624
Full (+ evidenced reflection)	**0.7744**	**0.6508**	**1.1641**	**2.6308**	**0.4051**	**0.5744**

Table 4: Ablation across the AnchorReviewer training trajectory. Each row adds one stage on top of the previous, and the last row replaces general-data reflection with reflection trained on AnchorReview-Anchored.

External benchmark. To verify that the gains are not specific to our test split, we further evaluate on the external DeepReview benchmark. AnchorReviewer continues to achieve the best Decision Accuracy/F1 and the highest Spearman / pairwise rating accuracy across both years, confirming the consistency of the observed advantages. The full table and analysis are provided in Appendix E.

5.3 Ablation and Analysis

Effect of training stages and reflection data. Table 4 ablates AnchorReviewer along its training trajectory. Substage-I (review warm-up on AnchorReview-General) lifts Decision Accuracy from 0.4821 (zero-shot Qwen3-VL-2B) to 0.6821, providing the basic review-format supervision the pretrained Qwen3-VL-2B itself lacks. Substage-II (query generation) further improves Accuracy and

F1 to 0.7283 / 0.5794, showing that learning to emit weakness queries is itself useful even before any retrieval.

Combining InRAG with reflection trained on *general* (non-anchored) data brings only minor Accuracy gains over Substage-II and drops F1 to 0.5512, showing unanchored reflection is unreliable on the rejection-class boundary. Training reflection on AnchorReview-Anchored data restores and outperforms Substage-II in F1 (0.7744 / 0.6508). This gap directly validates the value of the Anchored subset and evidence-grounded reflection design.

Multimodal evidence in InRAG matters. Table 6 ablates the modality of evidence fed into InRAG, with GPT-judge weakness/review scores. Removing InRAG entirely degrades every dimension, dropping Weakness Completeness from 2.92 to 2.13 and review-level Decision-Consistency from 2.74 to 2.55. Restricting InRAG to text-only chunks ("w/o Fig. in InRAG") recovers most of the gap on Weakness Overall and Review Overall, but Weakness Completeness still lags the full multimodal InRAG by 0.47, confirming that figure and table evidence is essential for thorough weakness analysis—precisely where a reviewer needs to point at a specific figure or table.

Why does UniEmb enable reliable multimodal retrieval? Table 5 compares CLIP, an off-the-shelf Qwen3-VL-Embedding (no fine-tuning), and our fine-tuned UniEmb on cross-modal evidence retrieval. CLIP collapses to R@1 = 1.2% and the original Qwen3-VL-Emb only reaches 15.0%, exposing a sizeable gap between paragraph-level pretraining and the fine-grained text–figure–table alignment required by our task. UniEmb doubles R@1 to 30.0% and improves MRR by 49% relative to the original encoder. Modality-decomposed retrieval results in Appendix (Tables 12,13,14,15) further show that the gap concentrates on *figure* chunks, where the off-the-shelf encoder collapses to 5% R@1 while UniEmb reaches 50%. This is consistent with a cross-modal alignment view: paragraph-level pretraining leaves a sizeable gap between text-form weakness queries and visual figure evidence, and UniEmb's contrastive training on aligned (query, multimodal-chunk) pairs closes this gap precisely where it is most severe.

5.4 Generalization Evaluation

To verify AnchorReviewer's gains stem from method design rather than venue exposure, we test it on held-out ICML'25 and NeurIPS'25 papers. None of these venues appear in the training data of CycleReviewer, DeepReviewer or AnchorReviewer. Across both venues, AnchorReviewer leads all competitors on every metric, outperforming the top fine-tuned baseline by roughly 18 D.Acc points and the best proprietary agent by 3–4 D.Acc points. Its ranking against baselines remains consistent across the two held-out sets, showing the coarse-to-fine, evidence-grounded design generalizes across review distributions even with training limited to ICLR. We note NeurIPS, ICML and ICLR share similar review conventions. True out-of-community evaluation (e.g., MLSys, OSDI) is hindered by unavailable public review data and left for future work. Full per-venue results are presented in Appendix D.

Method	R@1	R@3	R@5	R@10	MRR
CLIP	1.2	3.8	7.5	21.2	0.0949
Qwen3-VL-Emb	15.0	36.2	53.8	63.7	0.3045
UniEmb (Ours)	**30.0**	**47.5**	**66.3**	**83.8**	**0.4539**

Table 5: Performance comparisons of retrieval models.

Variant	Weaknesses				Review			
	Overall	Consi.	Acc.	Comple.	Overall	Consi.	Acc.	D. Consi.
w/o InRAG	2.32	1.83	2.85	2.13	3.38	3.11	3.68	2.55
w/o Fig. in InRAG	2.70	2.21	3.03	2.45	3.41	3.15	3.69	2.64
Full InRAG	**2.80**	**2.45**	**3.18**	**2.92**	**3.49**	**3.29**	**3.87**	**2.74**

Table 6: Ablations of components in InRAG.

6 Conclusion

In this work, we present the AnchorReviewer framework, enhancing the localization, credibility, and logical coherence of automated review generation. The Coarse-to-Fine Evidence-Grounded Review Generation Pipeline uses the RAG mechanism to extract evidence from multimodal chunks for detailed reviews. To tackle cross-modal retrieval, we introduce the Unified Embedding Model. The InRAG module enables query-triggered evidence retrieval and generation interaction, while the Reflection mechanism improves logical consistency. We also construct the annotated AnchorReview dataset for model training and evaluation. Through the Parallel Progressive Training Strategy, our model shows significant improvements in generation quality, evidence grounding, and logical coherence, outperforming traditional methods. We hope this work can inspire further advancements in the field of MLLM-based review generation and evidence-grounded review generation.

Limitations

The model's performance is largely contingent upon the quality of the input data, meaning that inaccuracies or biases in the training data could adversely affect the model's outputs. While our approach enhances the consistency and detail of the generated reviews, it may still encounter difficulties when addressing highly subjective or intricate academic writing that necessitates specialized domain knowledge. Furthermore, AnchorReviewer may face challenges when processing non-standard or poorly structured documents, as it excels primarily with well-structured academic papers. Moreover, the model's dependence on multimodal evidence makes PDF OCR capabilities a significant limiting factor. Current OCR technologies often struggle to process low-quality or complexly formatted PDFs, thereby reducing text extraction accuracy. Incorporating more accurate text input would markedly enhance the model's performance. Consequently, advancing PDF OCR capabilities is vital for the continued refinement of the model.

Additionally, the current iteration lacks a DeepResearch ability, limiting its ability to conduct deep web searches to verify whether the paper's ideas overlap with existing literature, which could raise concerns regarding originality. Furthermore, while the model's generation performance has markedly improved, integrating reinforcement learning could further refine its generative capabilities, enhancing its effectiveness in handling more complex review tasks.

Since peer review materials are publicly available solely for conference papers in artificial intelligence, our model can only conduct reviews on AI papers.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. 2026. Stop automating peer review without rigorous evaluation. *arXiv preprint arXiv:2605.03202*.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15651–15682, Suzhou, China. Association for Computational Linguistics.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. 2025. Vlm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8140–8158.

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

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. Nlpeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 5049–5073.

Charles W Fox, Jennifer Meyer, and Emilie Aimé. 2023. Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology*, 37(5):1144–1157.

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

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, pages 1647–1661.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. Disapere: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pages 1234–1249.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. 2025. Unveiling the merits and defects of llms in automatic re-

view generation for scientific papers. *arXiv preprint arXiv:2509.19326*.

Xuan Li, Yining Wang, Haocai Luo, Shengping Liu, Jerry Liang, Ying Fu, Jun Yu, Junnan Zhu, and 1 others. 2026. Bayesrag: Probabilistic mutual evidence corroboration for multimodal retrieval-augmented generation. *arXiv preprint arXiv:2601.07329*.

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

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025.

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*.

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. 2025. Agent reviewers: Domain-specific multimodal agents with shared memory for paper review. In *Forty-second International Conference on Machine Learning*.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and 1 others. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*.

Miryam Naddaf. 2025. Ai is transforming peer review—and many scientists are worried. *Nature*, 639(8056):852–854.

Vishisht Srihari Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B Shah. 2025. Detecting llm-generated peer reviews. *PLoS One*, 20(9):e0331871.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Automatically evaluating the paper reviewing capability of large language models. *arXiv preprint arXiv:2502.10886*.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17.

Pawin Taechoyotin and Daniel Acuna. 2025. Remor: Automated peer review generation with llm reasoning and multi-objective reinforcement learning. *arXiv preprint arXiv:2505.11718*.

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*.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *arXiv preprint arXiv:2504.09737*.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *International Conference on Learning Representations*, volume 2025, pages 3669–3709.

Shican Wu, Xiao Ma, Dehui Luo, Lulu Li, Xiangcheng Shi, Xin Chang, Xiaoyun Lin, Ran Luo, Chunlei Pei, and 1 others. 2025. Automated literature research and review-generation method based on large language models. *National Science Review*, 12(6):nwaf169.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. 2025. Universalrag: Retrieval-augmented generation over multiple corpora with diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pages 21074–21098.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Junpeng Yue, Xinrun Xu, Börje F Karlsson, and Zongqing Lu. 2025. Mllm as retriever: Interactively learning multimodal retrieval for embodied agents. In *International Conference on Learning Representations*, volume 2025, pages 31551–31580.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. 2025. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*.

Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, and 1 others. 2026. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment. *arXiv preprint arXiv:2601.01576*.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Chen, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu

Pu, Yuxuan Zhao, Zehua Xie, and 1 others. 2024. mr2ag: Multimodal retrieval-reflection-augmented generation for knowledge-based vqa. *arXiv preprint arXiv:2411.15041*.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. 2025a. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review. *arXiv preprint arXiv:2509.09912*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.

A Dataset Construction Details

ANCHORREVIEW is collected from all publicly accessible OpenReview submissions of ICLR (2018–2026), NeurIPS (2021–2025), and ICML 2025. All submission metadata and review contents are downloaded through the official OpenReview API. ICML appears only in 2025 because its review history on OpenReview only became publicly available that year; we still include it to broaden the venue coverage. The construction pipeline produces three hierarchically nested subsets, whose statistics we report from two complementary angles below: *per-venue composition* (Table 7) and *per-subset details* (Table 8).

Filtering pipeline. After (i) removing withdrawn or desk-rejected submissions, (ii) de-duplicating multi-version revisions and keeping only the initial submission, and (iii) discarding reviews shorter than 50 tokens, we retain 68,356 papers and 300,737 reviews (the GENERAL subset). Among these, 31.2K papers contain at least one locatable evidence reference (e.g., explicit "Table 3" or "Section 4.2" citation), from which layout parsing extracts 940.3K multimodal chunks (the EVIDENCED pool). Finally, atomic weakness decomposition on 28.7K of these papers produces 90K (query, evidence, weakness) triples (the ANCHORED subset).

Per-venue composition. Table 7 reports the venue-wise paper distribution. ICLR contributes the majority of the corpus (67.6%) thanks to its longest OpenReview coverage (2018–2026); NeurIPS contributes a substantial share (27.4%) covering 2021–2025; ICML adds a smaller portion (5.0%) that complements the other two venues with a different reviewing community despite its single-year window.

Per-subset details. Table 8 reports the dataset-level statistics of the three nested subsets. Items are denominated differently across subsets (reviews, chunks, triples) to reflect the supervision granularity each subset provides. Using the modality composition in §4, the 940.3K chunks in EVIDENCED correspond to approximately 494.6K text chunks, 199.3K equation chunks, 135.4K figure chunks, and 111.0K table chunks.

To improve the reliability of ANCHORREVIEW, we apply a three-level quality-control pipeline that combines automatic filtering, cross-model validation, and human verification. Automatic filtering operates at the corpus level and therefore benefits all three subsets (GENERAL, EVIDENCED, and ANCHORED), whereas cross-model validation and human verification are targeted at the fine-grained ANCHORED subset, which is the most error-sensitive component used for grounded weakness generation.

(i) Automatic filtering. This stage is applied at the *review* level and is therefore shared by all three subsets. We remove obviously noisy review instances, including cases with severe inconsistency between the reviewer rating and the final meta-review decision, and suspected machine-generated reviews detected by a perplexity-based detector together with stylometric indicators (e.g., abnormal repetition patterns and degenerate length-to-vocabulary ratios). Reviews shorter than 50 tokens are also excluded, as already noted in the filtering pipeline above. The filtered reviews then enter the downstream chunking and anchoring stages.

(ii) Cross-model validation. For a 5% random sample of the ANCHORED pool, we re-extract *(weakness, location)* pairs using an independent judge model (Claude Opus 4) under the same input protocol. We compute agreement on both weakness and location fields; samples with location-level Jaccard agreement below 0.6 are flagged for manual inspection. The resulting cross-model agreement is 98.6%.

(iii) Human verification. We further sample 300 triples for expert auditing. Three NLP graduate annotators assess each triple on three binary dimensions: *query validity*, *evidence relevance*, and *spatial accuracy*. Each sample is double-annotated, and disagreements are adjudicated by the third annotator. We follow the annotation guideline in §I. The overall validity is 92.4%, with Cohen's $\kappa = 0.81$, indicating substantial inter-annotator agreement.

Per-stage summary. Table 9 consolidates the scope, sample size, pass rate, and threshold of each quality-control stage. Sample sizes follow directly from the construction pipeline: stage (ii) re-annotates a 5% random sample of the 90K ANCHORED triples (i.e., $\sim$4.5K triples), and stage (iii) audits a fixed pool of 300 triples drawn from the same subset.

Venue	Period	#Papers	Share
ICLR	2018–2026	46,171	67.6%
NeurIPS	2021–2025	18,763	27.4%
ICML	2025	3,422	5.0%
Total		**68,356**	**100%**

Table 7: Per-venue composition of the ANCHORREVIEW corpus (GENERAL subset).

Subset	#Papers	#Items
GENERAL	68,356	300,737 reviews
EVIDENCED	31.2K	940.3K chunks
ANCHORED	28.7K	90K triples

Table 8: Per-subset statistics of ANCHORREVIEW.

A.1 Quality Control Details

B Evaluation Metrics Details

We used comprehensive metrics to evaluate the performance of our model.

To comprehensively evaluate the model's review generation capabilities, we designed several metrics covering decision accuracy, scoring precision, and subjective assessment.

For decision metrics, Accuracy measures overall correctness. The F1 score combines precision and recall.

For scoring metrics, Mean Absolute Error (MAE) calculates the average difference between predicted and true scores, Mean Squared Error (MSE) measures squared differences, and ± 1 accuracy measures the accuracy within a final score difference of 1 point.

For subjective evaluation, Consistency with Ground Truth (consis) checks review consistency, Overall Accuracy (acc) assesses review logic, Completeness and Detail (detail) evaluates thoroughness, and Contradictions checks for internal conflicts. Decision Consistency evaluates the alignment of final decisions. Additional metrics include Weakness Consistency, Weakness Accuracy, Weakness Detail, and Weakness Diversity, assessing the accuracy and multidimensional recognition of weakness evaluations.

C Training Details

This section gives the explicit training objectives referenced in §3; model sizes, optimiser, learning rates, and batch size are reported in §5. The overall stream/substage layout is depicted on the non-top-left portion of Figure 2 in the main text (the top-left portion shows the inference pipeline).

Data sources. The three streams are supervised on the three nested subsets of AnchorReview. Stream-I Substage-I (review warm-up) uses AnchorReview-General (paper–review pairs); Substage-II (query generation) and Substage-III (multi-task reflection) use AnchorReview-Anchored (query–evidence–weakness triples). Stream-II contrastively trains the unified encoder E_{uni} on (query, chunk) pairs sampled from AnchorReview-Evidenced, with hard negatives restricted to chunks from the same paper. Stream-III performs evidence-grounded review completion on ANCHORREVIEW-ANCHORED, using the current E_{uni} to retrieve in-context evidence for each query.

Shared NTP objective. Stream-I and Stream-III supervise language generation with next-token prediction (NTP):

$$\mathcal{L}_{\text{NTP}}(y \mid x) = -\sum_{t=1}^{|y|} \log P(y_t \mid y_{<t}, x), \qquad (5)$$

where x is the conditioning input and y the target token sequence.

Stream-I: Review Capability Training. The three sub-stages of Stream-I instantiate Eq. 5 with different conditioning inputs and targets.

Substage-I (Review warm-up).

$$\mathcal{L}_{\text{review-warmup}} = -\sum_{t=1}^{T} \log P(t \mid t_1, t_2, \ldots, t_{t-1}), \qquad (6)$$

where T is the total token count, t is the target token, and $t_1, t_2, \ldots, t_{t-1}$ are the input context tokens. The model learns to generate a complete review (summary, strength, weakness, and decision) from the input PDF.

Substage-II (Query generation).

$$\mathcal{L}_{\text{query-gen}} = -\sum_{t=1}^{T} \log P(t \mid q_1, q_2, \ldots, q_{t-1}), \qquad (7)$$

where q denotes the generated query tokens, and the objective maximises the likelihood of the query sequence.

Stage	Scope	Sample Size	Pass / Agree.	Threshold / Metric
(i) Automatic filtering	Raw → GENERAL	326,935 reviews	300,737 reviews	len $\geq$ 50 tok; PPL + stylometric
(ii) Cross-model validation	5% of ANCHORED	4,500 triples	98.6%	Jaccard $< 0.6 \Rightarrow$ human review
(iii) Human verification	Sample from ANCHORED	300 triples	92.4%	double-annotation; $\kappa = 0.81$

Table 9: Per-stage statistics of the ANCHORREVIEW quality-control pipeline.

Substage-III (Multi-task reflection).

$$\mathcal{L}_{\text{reflection}} = -\sum_{t=1}^{T} \log P(t \mid \text{summary}, \text{strength}, W, \text{previous review}). \tag{8}$$

The model consumes summary, strength, weakness W, and the previous review to produce a final decision that is internally consistent with the evidence-grounded weaknesses. The three sub-losses are summed as $\mathcal{L}_{\text{review}} = \mathcal{L}_{\text{review-warmup}} + \mathcal{L}_{\text{query-gen}} + \mathcal{L}_{\text{reflection}}$ and interleaved within the same epoch.

Stream-II: UniEmb Training. The unified encoder E_{uni} is trained with a contrastive InfoNCE objective:

$$\mathcal{L}_{\text{rag-r}} = -\log \frac{\exp(\text{sim}(q, c_{\text{true}})/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(q, c_k)/\tau)}, \tag{9}$$

where c_{true} is the gold evidence chunk for query q, $\{c_k\}$ includes hard negatives mined within the same paper, and τ is a temperature. Restricting negatives to the same paper forces E_{uni} to capture paper-internal layout cues (numbering, captions, neighbouring text) rather than topic-level cues; we found this to be essential for layout-grounded retrieval and verify it empirically in §5.

Stream-III: Review Completion Training. The weakness generator is supervised with

$$\mathcal{L}_{\text{rag-g}} = -\sum_{t=1}^{T} \log P(t \mid E_{\text{inline}}, q), \tag{10}$$

where E_{inline} denotes the retrieved relevant evidence and q is the corresponding query. For each anchored triple (q, e, w) from AnchorReview-Anchored, $E_{\text{inline}}(q)$ is computed under the *current* E_{uni}. Since E_{uni} is itself being updated by Stream-II, we periodically (every few thousand optimisation steps) refresh the cached top-k chunks for each query under the latest encoder, which prevents the generator from over-fitting to a stale retriever.

Progressive curriculum. The total objective $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{review}} + \lambda_2 \mathcal{L}_{\text{rag-r}} + \lambda_3 \mathcal{L}_{\text{rag-g}}$ is optimised under a progressive schedule on the weights λ_i. Concretely, training starts with λ_1 as the dominant weight so that the model first acquires the overall review form and the basic ability to emit queries. Once queries become non-trivial, λ_2 is phased in so that the retriever is trained on meaningful rather than random queries. Finally λ_3 is up-weighted to close the retrieve-and-generate loop. In our preliminary experiments, training all three streams from scratch with uniform weights led to noticeable degradation on both retrieval recall and decision accuracy.

D Out-of-Distribution Generalization

Table 10 reports rating-error (R. MSE, R. MAE), decision (D. Acc, D. F1), and ranking (R. Spearman, Pair. R. Acc) metrics on ICML'25 and NeurIPS'25 papers that are held out from training. Crucially, neither venue appears in the training corpora of any fine-tuned reviewer compared here: CycleReviewer is trained on ICLR'24 OpenReview reviews, DeepReviewer on ICLR'24 and ICLR'25 reviews, and our AnchorReviewer is trained only on the ICLR portion of AnchorReview. Prompt-based agent baselines (AgentReview, AI Scientist) do not undergo any review-specific fine-tuning.

AnchorReviewer-6B remains the best on every metric on both held-out venues, surpassing all prompt-based and fine-tuned baselines on rating-error, decision, and ranking simultaneously. The relative ordering of AnchorReviewer against all baselines is consistent with the in-domain comparison in Table 2, suggesting that the observed gains are method-driven rather than driven by venue-specific exposure. We further observe that fine-tuned baselines (CycleReviewer-8B and DeepReviewer-14B) suffer the largest accuracy drops relative to in-domain, while proprietary agent baselines (which are not fine-tuned on any review corpus) degrade less; AnchorReviewer remains robust on both axes despite training only on ICLR.

Method	Model	ICML'25						NeurIPS'25					
		R. MSE ↓	R. MAE ↓	D. Acc ↑	D. F1 ↑	R. Spearman ↑	Pair. R. Acc ↑	R. MSE ↓	R. MAE ↓	D. Acc ↑	D. F1 ↑	R. Spearman ↑	Pair. R. Acc ↑
AgentReview	Claude-Sonnet-4	3.6248	1.5421	0.6184	0.4621	0.1546	0.5421	3.5142	1.5187	0.6087	0.4548	0.1487	0.5358
	Gemini-2.5-Flash	3.4127	1.4856	0.6378	0.4823	0.1872	0.5612	3.3284	1.4632	0.6258	0.4734	0.1812	0.5547
AI Scientist	DeepSeek-V3.2	3.2851	1.4623	0.6527	0.5118	0.2487	0.5683	3.1894	1.4378	0.6428	0.5023	0.2403	0.5618
	GPT-4o-mini	3.5783	1.5278	0.6198	0.4612	0.1834	0.5524	3.4912	1.5021	0.6087	0.4524	0.1764	0.5462
	GPT-5.2	4.1247	1.6743	0.6353	0.4818	0.1567	0.5384	4.0234	1.6492	0.6258	0.4736	0.1502	0.5321
CycleReviewer	8B	5.6948	1.8237	0.5187	0.3201	0.1438	0.4592	5.5871	1.8056	0.5092	0.3162	0.1387	0.4534
DeepReviewer	14B	4.6248	1.6987	0.5092	0.2998	0.2348	0.5037	4.5132	1.6754	0.4982	0.2941	0.2287	0.4983
AnchorReviewer	6B	**2.9847**	**1.3526**	**0.6892**	**0.5712**	**0.3742**	**0.5894**	**2.8923**	**1.3274**	**0.6743**	**0.5587**	**0.3624**	**0.5821**

Table 10: Out-of-distribution evaluation on ICML'25 and NeurIPS'25. Both venues are held out from the training data of all fine-tuned reviewers (CycleReviewer, DeepReviewer, and our AnchorReviewer).

E External Benchmark Evaluation

To verify that the trends observed on our in-domain test split generalise across data sources, we further evaluate on the external DeepReview (Zhu et al., 2025b) benchmark, which provides held-out paper–review pairs from ICLR'24 and ICLR'25. Table 11 reports decision metrics (D. Acc, D. F1), rating-error metrics (R. MSE, R. MAE), and ranking metrics (R. Spearman, Pair. R. Acc) on both years.

AnchorReviewer-6B achieves the best Decision Accuracy/F1 and the best ranking-related metrics (Spearman and pairwise rating accuracy) on both ICLR'24 and ICLR'25, surpassing all prompt-based agent baselines (AgentReview, AI Scientist) and fine-tuned reviewers (CycleReviewer-8B/70B). The relative ordering of methods remains consistent with the in-domain results in Table 2, indicating that the gains of AnchorReviewer are not specific to our test split. On rating-error metrics, DeepReviewer-14B obtains the lowest R. MSE/MAE on this benchmark; we attribute this to the fact that DeepReviewer is also fine-tuned on ICLR review distributions and tends to produce ratings concentrated around the dataset mean, which lowers MSE/MAE but at the cost of decision and ranking quality, as reflected by its substantially lower D. Acc/F1 and Spearman.

F Ablation Studies on Multimodal RAG

We compared the retrieval results across different modalities, as shown in Tables 12, 14, 13, 15 . After training with our unified embedding approach, our model demonstrates exceptional performance in multimodal retrieval tasks. The results show that the model outperforms previous methods across multiple retrieval benchmarks, exhibiting substantial improvements in retrieval accuracy, especially in Section, Figure, and Table retrieval. This highlights the effectiveness of our unified embedding strategy in bridging different modalities and achieving superior retrieval results.

G More Qualitative Results

We provide additional cases that further demonstrate the model's exceptional performance. The reviews generated for these cases are accurate, with precise identification of weaknesses and well-supported evidence. The model consistently delivers detailed, logically coherent feedback, showcasing its ability to generate high-quality and reliable review decisions. These additional examples underline the robustness of the model in diverse scenarios, reaffirming its effectiveness in generating reviews that are both accurate and actionable.

H Annotation Prompt Template for GPT-5.1

We list below the prompt template used to decompose a detailed review bullet into a structured (*query*, *location reference*) pair. Curly braces denote slots filled at runtime.

```
You are a meta-reviewer assistant.
Given a single weakness bullet from a
peer review, extract:
(1) query: a self-contained question
that captures what the reviewer wants
the authors to address;
(2) location: the specific element(s)
of the paper that the reviewer is
referring to. Use the canonical
form <Type>:<Number>[:Sub], e.g.
Table:3, Figure:2:bottom, Section:4.2,
Equation:5. Use None if no concrete
location is mentioned.
Return strict JSON: {"query": ...,
"location": [...]}.
Bullet: {{bullet}}
Paper outline (sections, figures,
tables): {{outline}}
```

I Human Evaluation Guideline

We sample 300 triples from AnchorReview-Anchored and ask three NLP graduate students

Method	Model	ICLR'24						ICLR'25					
		R. MSE ↓	R. MAE ↓	D. Acc ↑	D. F1 ↑	R. Spearman ↑	Pair. R. Acc ↑	R. MSE ↓	R. MAE ↓	D. Acc ↑	D. F1 ↑	R. Spearman ↑	Pair. R. Acc ↑
AgentReview	Claude-3.5	2.8878	1.2715	0.4333	0.3937	0.1564	0.5526	2.8406	1.2989	0.2826	0.2541	-0.0219	0.5432
	Gemini-2.0	3.1943	1.3418	0.44	0.4318	-0.0252	0.5044	2.6186	1.217	0.4242	0.4242	0.0968	0.5496
	DeepSeek-V3	1.9479	1.0735	0.4105	0.3403	0.3542	0.6096	1.9951	1.1017	0.314	0.2506	0.1197	0.5702
AI Scientist	GPT-o1	4.3414	1.7294	0.45	0.4424	0.2621	0.5881	4.3072	1.7917	0.4167	0.4157	0.2991	0.6318
	Claude-3.5	3.4447	1.5037	0.4787	0.4513	0.0366	0.5305	3.0992	1.35	0.5579	0.444	-0.0219	0.5169
	Gemini-2.0	4.9297	1.8711	0.5743	0.5197	0.0745	0.5343	3.9232	1.647	0.6139	0.4808	0.2565	0.604
	DeepSeek-V3	4.7337	1.7888	0.56	0.5484	0.231	0.5844	4.8006	1.8403	0.4059	0.3988	0.0778	0.5473
	Deepseek-R1	4.1648	1.6526	0.5248	0.4988	0.3256	0.6206	4.7719	1.8099	0.4259	0.4161	0.3237	0.6289
CycleReviewer	8B	2.8911	1.2371	0.6353	0.5528	0.2801	0.5993	2.4461	1.2063	0.678	0.5586	0.2786	0.596
	70B	2.487	1.2514	0.6304	0.5696	0.3356	0.616	2.4294	1.2128	0.6782	0.5737	0.2674	0.5928
DeepReviewer	14B	**1.3137**	**0.9102**	0.6406	0.6307	0.3559	0.6242	**1.3410**	**0.9243**	0.6878	0.6227	0.4047	0.6402
AnchorReviewer	6B	2.3145	1.0852	**0.7635**	**0.7258**	**0.5284**	**0.6612**	2.2702	1.0710	**0.7774**	**0.7406**	**0.5408**	**0.6723**

Table 11: External-benchmark comparison of review quality on the DeepReview ICLR'24 / ICLR'25 test sets. For Gemini-2.0, we use Gemini-2.0-Flash-Thinking.

Table 12: Section Retrieval Results.

Model	R@1	R@3	R@5	R@10	MRR
CLIP	5.0%	15.0%	30.0%	75.0%	0.1999
Original Qwen3-VL-Emb	20.0%	45.0%	70.0%	85.0%	0.4143
UniEmb (ours)	35.0%	50.0%	85.0%	100.0%	0.5255

Table 13: Figure Retrieval Results.

Model	R@1	R@3	R@5	R@10	MRR
CLIP	0.0%	0.0%	0.0%	0.0%	0.0539
Original Qwen3-VL-Emb	5.0%	25.0%	35.0%	50.0%	0.1952
UniEmb (ours)	50.0%	65.0%	75.0%	90.0%	0.6215

Table 14: Table Retrieval Results.

Model	R@1	R@3	R@5	R@10	MRR
CLIP	0.0%	0.0%	0.0%	5.0%	0.0649
Original Qwen3-VL-Emb	20.0%	35.0%	60.0%	70.0%	0.3453
UniEmb (ours)	20.0%	45.0%	65.0%	85.0%	0.3890

Table 15: Equation Retrieval Results.

Model	R@1	R@3	R@5	R@10	MRR
CLIP	0.0%	0.0%	0.0%	5.0%	0.0609
Original Qwen3-VL-Emb	15.0%	30.0%	40.0%	50.0%	0.2633
UniEmb (ours)	15.0%	30.0%	40.0%	60.0%	0.2795

to judge them on three binary criteria. **(i) Query Validity (Y/N)**: does the extracted query faithfully capture the concern expressed in the original review bullet? **(ii) Evidence Relevance (Y/N)**: is the chunk identified by the bounding box actually the piece of evidence the reviewer is criticizing? **(iii) Spatial Accuracy (Y/N)**: does the bounding box tightly cover the referenced element, without including unrelated regions or omitting the key subarea (e.g., a specific row of a table)? Each triple is double-annotated; disagreements are resolved by a third annotator. We report per-criterion accuracy, overall validity (all three criteria Y), and Cohen's κ in §4.

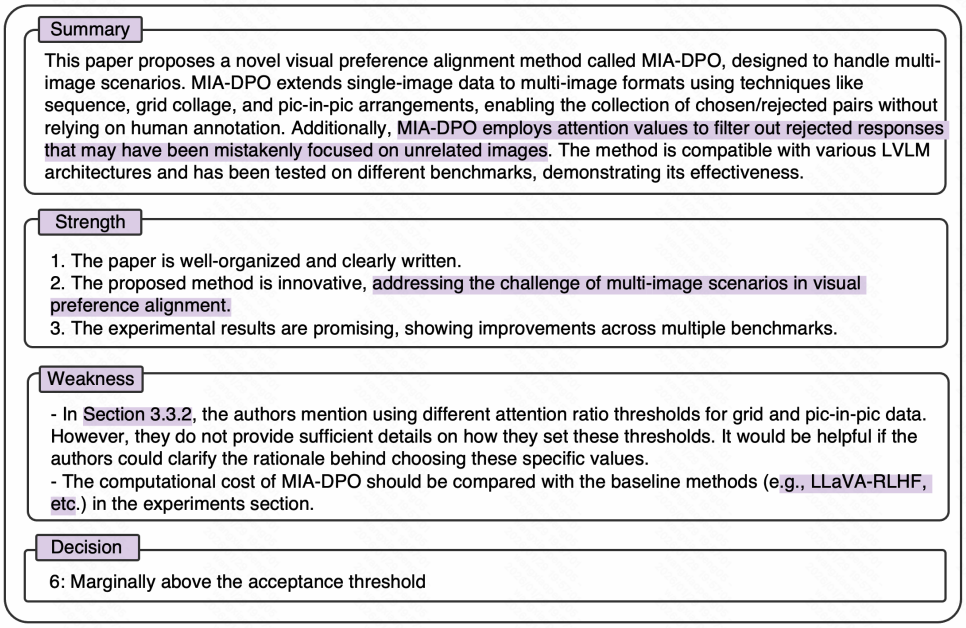

Figure 4: More generated review cases by AnchorReviewer.

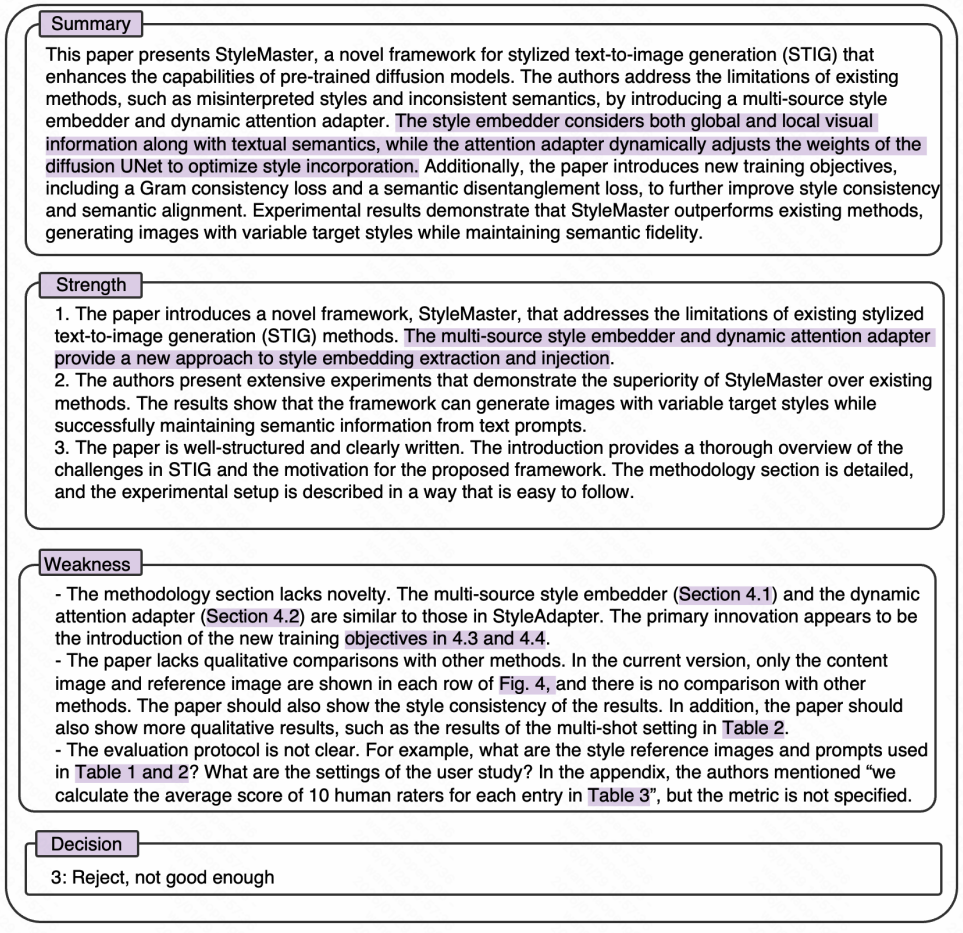

Figure 5: More generated review cases by AnchorReviewer.

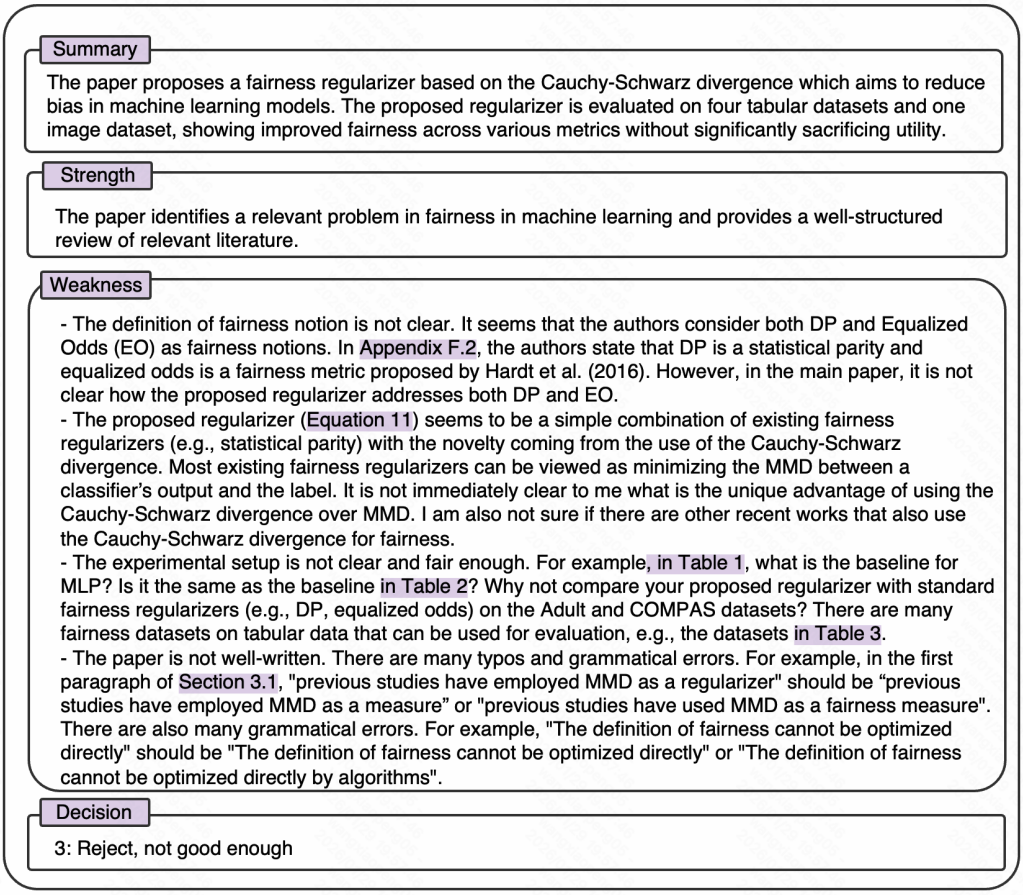

Figure 6: More generated review cases by AnchorReviewer.

Summary

This paper introduces a framework for quantifying deception in interactions, based on the concept of regret. The authors propose that regret provides a way to measure how deceptive an agent's actions are perceived by another agent. They formalize this mathematically within a partially observable Markov decision process (POMDP) model, where deception is quantified as the difference in listener rewards when acting according to deceptive versus truthful strategies. The framework is evaluated through human studies and applied to dialogue systems, demonstrating that the proposed regret-based metric aligns with human perceptions of deception.

Strength

- The paper is well-organized, with a clear structure that guides the reader through the theoretical formulation, experimental setup, and results.
- The POMDP-based framework is innovative, offering a structured way to quantify deception that takes into account both the listener's beliefs and the expected utility.
- The paper includes a comprehensive evaluation, with human studies and comparisons to dialogue systems. This multi-faceted approach strengthens the credibility of the proposed framework.

Weakness

- Definition 2.1 is not clear. What is L's policy? It is represented by a function from observations to actions, but this is not clear. Also, the observation function o_L is not clearly defined. The notation is also not strong enough. It should be defined that the notation o_L(s) means the observation function. The definition of the POMDP in Definition 2.2 is also not well justified. What is the POMDP? Is it a Markov chain? It is not explained well. Also, the reward function R is not defined. It is only mentioned in the paragraph of the POMDP. What is the role of R in the definition of the POMDP?
- 1. **The paper lacks a clear rationale for using regret to quantify deception.** In Section 2.3, the paper introduces the concept of regret, defined as the difference in listener rewards under deceptive actions versus truthful actions (Equation 1). However, the paper does not adequately explain the rationale behind this choice. Why is it appropriate to use regret as a measure of deception? Are there alternative metrics that could better capture this aspect?
- The paper is limited in the scope of the scenarios. The scenarios are based on simple binary features, and the paper does not discuss how to apply the approach to more complex real-world interactions. The scenarios are also not well justified, and it is not clear to me how they are relevant to human deception. The three scenarios considered in the paper are listed in Appendix E.
- In Section 2.4, the authors claim that one of the main reasons listeners do not deceive is because they do not think they are being deceived. I do not think this is a reasonable assumption to make. Why would L not think they are being deceived? There is no evidence that this is the case, and it is not clear to me why it would be reasonable to assume this.
- The paper does not discuss limitations and assumptions, which are important to understand when the framework can be used and when it cannot be used. For example, in the case of the negotiation task, it is not clear when the framework will fail. The authors should provide more discussion on this point. Similarly, the statement "The framework is only valid when the speaker's goal is to deceive the listener" (line 200) needs more clarification. It is not clear in which scenarios the framework can be used and in which scenarios it cannot be used. The authors should provide more discussion on this point.

Decision

3: Reject, not good enough

Figure 7: More generated review cases by AnchorReviewer.

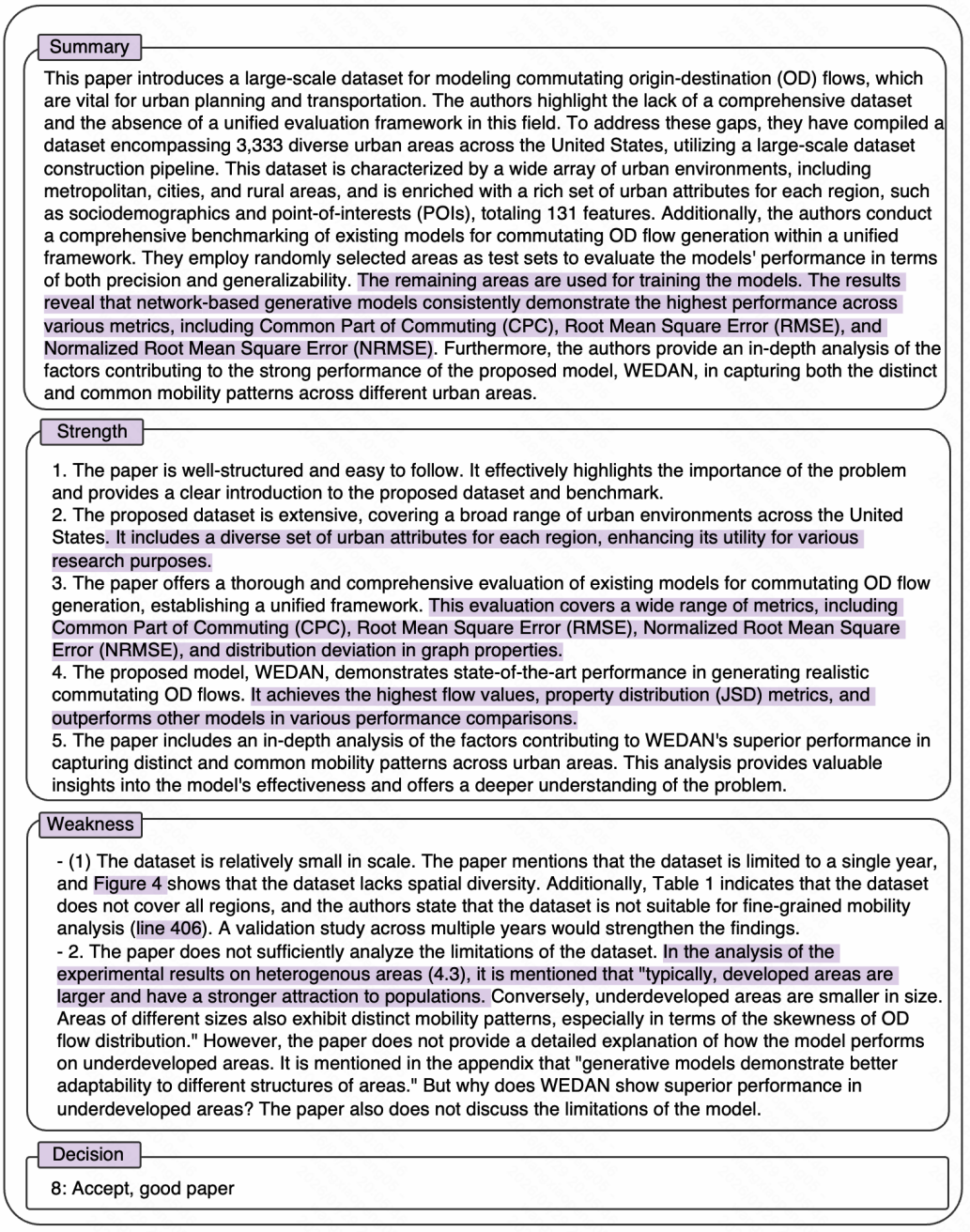

Figure 8: More generated review cases by AnchorReviewer.