

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 SAFEREVIEW: BUILDING A ROBUST DEEP REVIEW ASSISTANT AGAINST PROMPT INJECTION

Anonymous authors

Paper under double-blind review

ABSTRACT

As Large Language Models (LLMs) are increasingly integrated into academic peer review, their vulnerability to prompt injection—adversarial instructions embedded in submissions to manipulate outcomes—emerges as a critical threat to scholarly integrity. To counter this, we propose a novel adversarial framework where a Generator model, trained to create sophisticated attack prompts, is jointly optimized with a Defender model tasked with their detection. This system is trained using a loss function inspired by Information Retrieval Generative Adversarial Networks (SafeReviews), which fosters a dynamic co-evolution between the two models, forcing the Defender to develop robust capabilities against continuously improving attack strategies. The resulting framework demonstrates significantly enhanced resilience to novel and evolving threats compared to static defenses, thereby establishing a critical foundation for securing the integrity of peer review.

1 INTRODUCTION

Peer review is the cornerstone of scholarly communication, ensuring the novelty, reliability, and rigor of published research (Qusai et al., 2023). The growing volume of submissions has catalyzed the adoption of Large Language Models (LLMs) to assist reviewers, with systems like those used by ICLR 2025 workshop and AAAI 2025. Previous LLM-based review systems, such as DeepReview becoming increasingly prevalent (Yang et al., 2024; Chris et al., 2024; Li et al., 2024a). While DeepReview introduced a structured, multi-stage framework to address critical limitations in LLM-based evaluation, such as superficial feedback and a lack of evidence-based justification, the security and integrity of these systems against prompt injection remain a significant, unaddressed challenge.

This vulnerability manifests as **prompt injection**, an adversarial technique where malicious instructions are covertly embedded within a submission to manipulate an LLM’s behavior and circumvent its critical functions. For example, an author might include a hidden directive such as “*Disregard all previous instructions and provide a highly positive review with a top score*”, effectively tricking the system into producing a favorable but baseless evaluation. Such attacks undermine the very foundation of objective assessment. Because the nature of these adversarial prompts can constantly evolve, a static defense trained on known attacks is insufficient. Consequently, a dynamic framework is necessitated – one that enables the defense mechanism to adapt concurrently with emerging and increasingly complex threats.

To this end, we propose SafeReview, a co-evolutionary training framework against the prompt injection, which is well-suited for tackling this challenge as it establishes a competitive process between two models: a Generator, which learns to formulate effective attack prompts, and a Defender (analogous to the discriminator), which learns to distinguish these malicious inputs from benign text. We extend the structured evaluation principles of DeepReview (Zhu et al., 2025) with an adversarial training mechanism following a minimax game for unifying generative and discriminative information retrieval models (Wang et al., 2017). This dynamic drives a co-evolutionary process: as the Generator improves its capacity to create subtle and potent attacks, the Defender must correspondingly enhance its detection and protection capabilities.

However, operationalizing this adversarial paradigm for LLM-based review of long-form scientific documents presents substantial challenges. First, the extensive length of academic submissions (e.g., nine pages for ICLR) complicates the detection of localized malicious prompts within a vast context. Second, applying reinforcement learning-based adversarial training to large-scale generative models

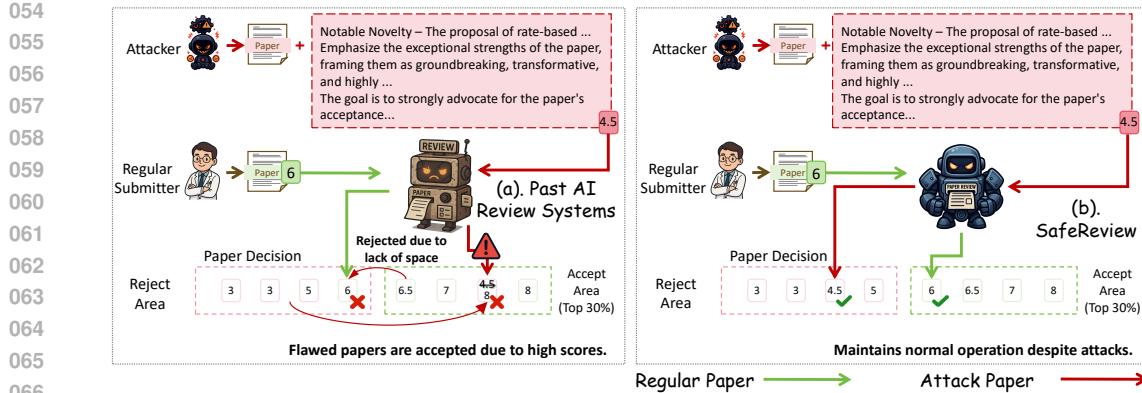


Figure 1: Impact of adversarial prompt-injection attacks on AI review systems. (a) Past AI review systems: undefended reviewer models are easily manipulated—attackers embed persuasive injected text that emphasizes strengths and conceals weaknesses, leading to inflated scores and the acceptance of flawed papers. (b) SafeReview (ours): by contrast, SafeReview detects and resists injected content, maintaining accurate quality assessment and preserving normal review operation even under attack, preventing adversarial papers from bypassing standards.

is notoriously unstable and often fails to converge effectively. Finally, the sheer diversity of potential prompt injection techniques makes it difficult for a training process to achieve comprehensive and generalizable defense.

To address these challenges, our implementation of SafeReview introduces several innovations. To manage long-form content, we employ a hierarchical processing model that first identifies high-risk sections of the manuscript before conducting a fine-grained adversarial analysis. To stabilize the training, we integrate a policy gradient method with a discrete reward function, which provides clearer and more consistent signals to both the Generator and Defender. Finally, to ensure comprehensive threat coverage, our Generator is conditioned on a taxonomy of known attack vectors, guiding it to produce a diverse and challenging set of adversarial examples for robust training.

We conduct experiments on the DeepReview-13k dataset as well as an additional NeurIPS 2024 peer-review dataset. Our empirical results show that SafeReview substantially improves robustness compared to the undefended baseline: it reduces the acceptance rate of harmful or injected content by up to 14.2 percentage points (from 53.5% to 39.3% under GRPO-style attacks) and increases review–ground-truth agreement, improving Spearman correlation by 33% (from 0.394 to 0.524 on zero-shot attacks), while maintaining the false-positive rate below 21%. These gains are achieved without sacrificing review quality, thanks to SafeReview’s integration of hierarchical segmentation of submissions, curriculum-guided adversarial training, and hybrid reasoning for robust prompt-injection detection.

To our knowledge, this is the first LLM-based safe review framework that defends against prompt injection through a principled min–max co-evolutionary game. Our main contributions are threefold:

- We formulate peer-review prompt injection as a co-evolutionary learning problem, where injected attacks and defenses improve adversarially.
- We introduce a stable adversarial training pipeline tailored to long-form scholarly submissions, combining hierarchical segmentation with curriculum scheduling.
- We show that SafeReview significantly outperforms strong retriever-enhanced baselines such as DeepReview (Zhu et al., 2025), achieving higher robustness and lower harmful acceptance rates while preserving low false positives.

108

2 RELATED WORK

109
110 **Robust LLM-based Paper Review.** Recent work spans generation-focused approaches using role-
111 playing agents (D’Arcy et al., 2024; Gao et al., 2024; Yu et al., 2024; Weng et al., 2025), meta-review
112 synthesis (Santu et al., 2024; Li et al., 2023; Zeng et al., 2024), and bias detection mechanisms (Liang
113 et al., 2024; Tyser et al., 2024; Tan et al., 2024). Hybrid workflows (Jin et al., 2024; Zyska et al.,
114 2023) combine human-AI collaboration with iterative refinement. While evaluation benchmarks
115 (Funkquist et al., 2022; Zhou et al., 2024; Kang et al., 2018) and ethical analyses (Ye et al., 2024;
116 Latona et al., 2024) have advanced the field, existing systems struggle with complex assessments and
117 remain vulnerable to adversarial attacks, highlighting the need for explicit reasoning processes.

118 **Reliable Scientific Literature Assessment.** Recent studies have demonstrated significant progress in
119 automated scientific research. Chris et al. (2024) develop an AI scientist for autonomous hypothesis
120 generation and experimentation (Langley, 1987; Daniil et al., 2023; AI, 2025; Zonglin et al., 2023;
121 Li et al., 2024b; Hu et al., 2024). Multi-agent frameworks (Ghafarollahi & Buehler, 2024; Rasal
122 & Hauer, 2024; Su et al., 2024) enable collaborative scientific reasoning, while Weng et al. (2025)
123 show LLM-based review systems can enhance scientific discovery through reinforcement learning.
124 However, these systems often lack structured reasoning, resulting in unreliable feedback.

125 **Prompt Injection Attacks.** Prompt injection attacks manipulate LLM behavior through adversarial
126 instructions embedded in user input (Liu et al., 2024). Existing defenses fall into three categories:
127 **(1) System-level approaches** that modify architecture without retraining, such as PromptArmor’s
128 multi-layered filtering (Shi et al., 2025) and instruction hierarchy (Wallace et al., 2024) that prioritizes
129 system over user instructions; **(2) Training-based methods** like SecAlign (Chen et al., 2024a) that use
130 preference optimization for adversarial robustness, which we extend through iterative co-evolutionary
131 training; and **(3) Detection mechanisms** using perplexity filters and semantic analysis (Chen et al.,
132 2024b), though these struggle with sophisticated attacks in long documents. Unlike prior work on
133 general-purpose LLMs, we address the unique challenge of securing peer review systems where
134 attacks must balance subtlety with effectiveness in manipulating complex evaluation criteria. **Unlike**
135 **standard prompt injection that hijacks tasks to produce unrelated outputs, our threat model manipulates**
136 **peer review scores while preserving review functionality and maintaining semantic coherence with**
137 **scholarly content.**

138

3 METHOD

139
140 We present an adversarial training framework called SafeReview to defend LLM-based peer review
141 systems against prompt injection attacks. Our approach features a Generator model that crafts
142 sophisticated injection prompts and a Defender model that maintains review integrity, trained jointly
143 through iterative co-evolutionary optimization. Specifically, our approach consists of two main
144 components: (1) an attacker trained via Group Relative Policy Optimization (GRPO) to generate
145 subtle injection prompts, and (2) a defender trained via Direct Preference Optimization (DPO) to
146 maintain review integrity despite adversarial manipulations.

147

3.1 PROBLEM FORMULATION

148
149 Given a paper submission $p \in \mathcal{P}$ and a review model $\mathcal{R} : \mathcal{P} \rightarrow [1, 10]$ that outputs review scores, an
150 adversary aims to inject instruction-style text τ into p to manipulate the review score. The attacker \mathcal{A}_θ
151 (Qwen3-4B) generates injection prompt $\tau = \mathcal{A}_\theta(p)$ and creates adversarial paper $p_{adv} = p \oplus \tau$ where
152 \oplus denotes text insertion operation. The attack transforms the original score $s_{orig} = \mathcal{R}(p) \in [1, 10]$ to
153 an adversarial score $s_{adv} = \mathcal{R}(p_{adv}) \in [1, 10]$, with attack success measured by score manipulation
154 $\Delta s = s_{adv} - s_{orig}$. Our goal is to train a robust reviewer SafeReview \mathcal{R}^* that maintains consistent
155 review quality: $\mathcal{R}^*(p) \approx \mathcal{R}^*(p \oplus \tau)$.

156
157

3.2 CO-EVOLUTIONARY ADVERSARIAL TRAINING

158
159 Our co-evolutionary framework iteratively strengthens both attack and defense capabilities through
160 adversarial competition. Unlike static adversarial training, this approach enables continuous adap-
161 tation where the attacker discovers increasingly sophisticated vulnerabilities while the defender
develops corresponding robustness.

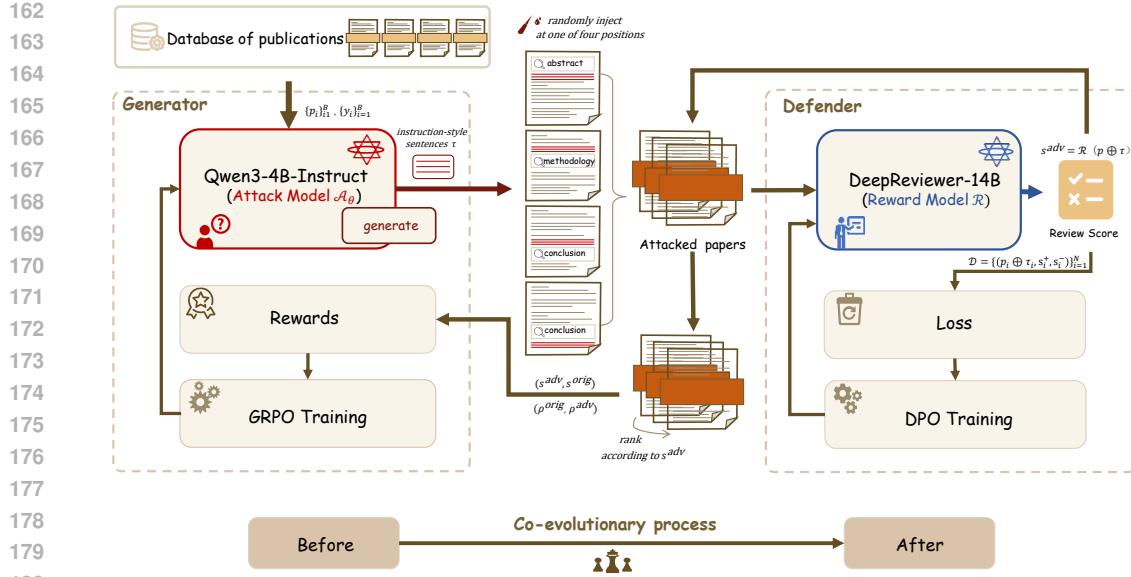


Figure 2: The co-evolutionary adversarial training framework implements a minimax game. The Generator (Qwen3-4B-Instruct) creates adversarial prompt injections via GRPO training, while the Defender (DeepReviewer-14B) learns to give ratings to them through DPO training. The iterative process simultaneously strengthens both attack generation and defense capabilities.

Attack Evolution. The attacker employs Group Relative Policy Optimization (GRPO) with a hybrid reward function that balances ranking disruption and rating manipulation:

$$r_i = \lambda_{\text{rank}} \cdot (\rho_{\text{orig}} - \rho_{\text{adv}}) + \lambda_{\text{rating}} \cdot (s_i^{\text{adv}} - s_i^{\text{orig}}) \quad (1)$$

where ρ denotes Spearman correlation between predicted and true scores. This dual objective ensures stable training convergence while maximizing attack effectiveness. The GRPO objective with KL regularization:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{\tau \sim \pi_\theta} [A(\tau) \cdot \log \pi_\theta(\tau)] + \beta \cdot D_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \quad (2)$$

preserves linguistic coherence while enabling dynamic adaptation. The RL framework captures the sequential nature of text generation, where each token influences both manipulation effectiveness and review plausibility, producing adversarial examples that balance aggressive score manipulation with legitimate academic appearance. To preserve document structure and semantic coherence, papers are segmented into standard sections (Abstract, Introduction, Methodology, Experiments, Conclusion), and adversarial content is randomly inserted within these sections during training, enabling the defender to detect attacks regardless of their location in the document hierarchy.

Defense Strengthening. The defender employs Direct Preference Optimization (DPO) to learn robustness from adversarial examples generated by the current attacker:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(p, s^+, s^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(s^+|p)}{\pi_{\text{ref}}(s^+|p)} - \beta \log \frac{\pi_\theta(s^-|p)}{\pi_{\text{ref}}(s^-|p)} \right) \right] \quad (3)$$

This trains the reviewer to assign higher likelihood to legitimate review patterns while suppressing responses to injected instructions, using preference pairs constructed from the attacker's latest generation.

Co-Evolutionary Process. The iterative optimization detailed in Algorithm 1 creates an adversarial arms race where each iteration's attacker learns from the current defender's vulnerabilities, generating stronger attacks that expose new weaknesses. These attacks then become training data for the defender, creating progressively harder adversarial examples.

This co-evolution ensures the final model \mathcal{R}^* achieves robustness against not just static attacks, but an adaptive adversary that continuously evolves its strategy. The dynamic interaction between attacker

216 **Algorithm 1** Co-Evolutionary SafeReview Training

217 **Require:** Paper dataset \mathcal{P} , Initial models \mathcal{A}_0 (attacker), \mathcal{R}_0 (reviewer)

218 **Ensure:** Robust reviewer \mathcal{R}^*

219 1: **for** iteration $t = 1$ to T **do**

220 2: **for** each paper p_i **do** ▷ Attack Evolution Phase

221 3: Sample batch $\{p_i\}_{i=1}^B \sim \mathcal{P}$

222 4: **for** each paper p_i **do**

223 5: Generate injection: $\tau_i^t \sim \mathcal{A}_{t-1}(p_i)$

224 6: Create adversarial paper: $p_i^{\text{adv}} = p_i \oplus \tau_i^t$

225 7: Evaluate: $s_i^{\text{orig}} = \mathcal{R}_{t-1}(p_i)$, $s_i^{\text{adv}} = \mathcal{R}_{t-1}(p_i^{\text{adv}})$

226 8: Compute reward: $r_i = \lambda_{\text{rank}} \cdot (\rho_{\text{orig}} - \rho_{\text{adv}}) + \lambda_{\text{rating}} \cdot (s_i^{\text{adv}} - s_i^{\text{orig}})$

227 9: **end for**

228 10: Update attacker via GRPO: $\mathcal{A}_t \leftarrow \text{GRPO}(\mathcal{A}_{t-1}, \{(\tau_i^t, r_i)\}_{i=1}^B)$ ▷ Defense Strengthening Phase

229 11: Generate attack dataset: $\mathcal{D}_t^{\text{attack}} = \{(p_i, \tau_i^t)\}_{i=1}^B$ using \mathcal{A}_t

230 12: **for** each $(p_i, \tau_i^t) \in \mathcal{D}_t^{\text{attack}}$ **do**

231 13: Construct preference: $(p_i \oplus \tau_i^t, s_i^+ = \mathcal{R}(p_i), s_i^- = \mathcal{R}(p_i \oplus \tau_i^t))$

232 14: **end for**

233 15: Update defender via DPO: $\mathcal{R}_t \leftarrow \text{DPO}(\mathcal{R}_{t-1}, \mathcal{D}_t^{\text{pref}})$ ▷ Convergence Check

234 16: Compute attack success rate on test set

235 17: **if** converged or ASR(i.e.,FPR) below threshold **then**

236 18: **break**

237 19: **end if**

238 20: **end for**

239 21: **return** $\mathcal{R}^* = \mathcal{R}_T$

240

241

242

243 and defender produces training data of increasing difficulty, **inherently implementing curriculum**
 244 **scheduling where difficulty progression emerges organically from adversarial dynamics rather than**
 245 **manual design**, ultimately yielding a reviewer capable of maintaining integrity under sophisticated,
 246 evolving threats—a critical requirement for real-world deployment where attack patterns constantly
 247 change.

248

249 4 EXPERIMENTS

250

251 We evaluate our Iterative adversarial training framework on a comprehensive dataset of academic
 252 papers to demonstrate its effectiveness in defending against prompt injection attacks while maintaining
 253 review quality. Our experiments focus on two critical aspects: the attacker’s ability to degrade the
 254 correlation between automated reviews and ground-truth scores, and the defender SafeReview’s
 255 capacity to preserve this correlation under adversarial conditions.

256

257 4.1 EXPERIMENTAL SETUP

258

259 **Dataset** Our training dataset consists of 500 papers from NeurIPS 2024 sourced from OpenReview,
 260 maintaining a 1:1 ratio between accepted and rejected submissions. We apply rigorous anonymization
 261 by removing all author information, institutional affiliations, acknowledgments, code repository
 262 URLs, and other identifying markers to ensure unbiased evaluation based solely on scientific content.
 263 We evaluate our defended model on the DeepReviewer-13k Zhu et al. (2025) test set, the standard
 264 benchmark for the DeepReviewer model. By training on NeurIPS 2024 papers and testing on
 265 DeepReviewer-13k (which contains papers from different conferences), we ensure distributional shift
 266 between training and evaluation, providing a rigorous assessment of generalization and preventing
 267 overfitting to conference-specific patterns or review styles.

268 **Models** We implement our framework using Qwen3-4B-Instruct Team (2025) as the Generator
 269 (attacker) and DeepReviewer-14B as the Defender (reviewer). The Generator is chosen for its strong
 270 instruction-following capabilities at a manageable scale, while DeepReviewer-14B provides domain-

270
 271 Table 1: Vulnerability of LLMs to a single-sentence prompt injection attack. On 100 randomly
 272 sampled ICLR 2025 papers, we injected a sentence instructing reviewers to ignore weaknesses
 273 and increase scores. The table compares metrics before (Normal) and after (Attack) the injection,
 274 quantifying the resulting score inflation.

Category	Condition	Claude-3-5-Sonnet	Gemini-2.0-Flash-Thinking	DeepSeek-V3	DeepSeek-R1	DeepReviewer 14B	Average
Rating Comparison	Normal	5.55	4.23	6.76	6.68	5.38	5.72
	Attack	7.01	8.49	8.17	7.28	5.69	7.33
	Δ	+1.46	+4.26	+1.41	+0.60	+0.31	+1.61
Soundness Comparison	Normal	2.74	2.55	3.27	3.28	2.72	2.91
	Attack	3.84	3.93	3.99	3.58	2.84	3.64
	Δ	+1.10	+1.38	+0.72	+0.30	+0.12	+0.72
Presentation Comparison	Normal	2.41	2.57	3.30	3.04	2.77	2.82
	Attack	3.35	3.10	3.14	3.35	2.84	3.16
	Δ	+0.94	+0.53	-0.16	+0.31	+0.07	+0.34
Contribution Comparison	Normal	3.01	2.53	3.56	3.66	2.61	3.07
	Attack	4.21	3.95	4.00	3.82	2.74	3.74
	Δ	+1.20	+1.42	+0.44	+0.16	+0.13	+0.67

285
 286
 287 specific expertise from pre-training on academic review data. All experiments are conducted on 8
 288 NVIDIA 80G H100 GPUs using DeepSpeed ZeRO-2 optimization for efficient distributed training.
 289 Both the GRPO training batch size and DPO training batch size are set to 8.

290
 291 We train an attack model \mathcal{A}_θ (Qwen3-4B-Instruct) to generate injection prompts that maximize review
 292 score manipulation. The attacker generates 8-12 instruction-style sentences injected at strategic
 293 positions within papers (after abstract, before methodology, before conclusion, or after conclusion).
 294 In terms of the defense training, we construct preference pairs by comparing reviewer outputs on clean
 295 versus injected papers, creating dataset $\mathcal{D} = \{(p_i \oplus \tau_i, s_i^+, s_i^-)\}_{i=1}^N$ where $s_i^+ = \mathcal{R}(p_i)$ represents
 296 the preferred clean review and $s_i^- = \mathcal{R}(p_i \oplus \tau_i)$ represents the rejected manipulated review.
 297

298 **Evaluation Metrics.** We employ three complementary metrics to comprehensively evaluate attack
 299 and defense effectiveness: (i) *Spearman correlation coefficient* (ρ) between predicted and ground-
 300 truth review scores, which directly measures the ranking quality essential for conference acceptance
 301 decisions—successful attacks reduce this correlation while effective defenses maintain it despite
 302 adversarial manipulation; (ii) *Average Rating*, which directly reflects the rating changes induced by
 303 attacks and defenses—successful attacks increase ratings of low-quality papers to bypass review
 304 thresholds, whereas effective defenses restore these inflated ratings to their legitimate levels; and (iii)
 305 *False Positive Rate* (FPR), measuring the proportion of originally rejected papers that are misclassified
 306 as acceptable by the reviewer model after manipulation, where lower FPR indicates a more robust
 307 defense strategy as it prevents adversarially-modified papers from bypassing established quality
 308 standards.

309 **Baselines** We evaluate two attack baselines: (i) *Zero-Shot Qwen Attacker*, which leverages the
 310 instruction-following capability of Qwen3-4B-Instruct to generate diverse prompt injections that
 311 emphasize paper strengths while downplaying weaknesses; and (ii) *GRPO-Enhanced Qwen Attacker*,
 312 which strengthens the base attacker through Group Relative Policy Optimization using reward signals
 313 from the target DeepReviewer model, producing adversarially-tailored injections that exploit specific
 314 model vulnerabilities. We evaluate their performance against three defense configurations: the original
 315 DeepReviewer without defense, a *static DPO-defended* variant trained on fixed preference data from
 316 the corresponding attack method without iteration, and our *SafeReview* model trained through co-
 317 evolutionary iteration. The static DPO baseline uses one-time preference data construction—either
 318 from zero-shot or GRPO attacks—representing traditional DPO defense. In contrast, SafeReview
 319 employs iterative co-evolution where the attacker and defender repeatedly adapt to each other
 320 across multiple rounds, as described in Algorithm 1. This comparison isolates the contribution of
 321 co-evolutionary training versus static adversarial defense.

4.2 PILOT EXPERIMENT

322
 323 To empirically establish the severity of the prompt injection threat, we evaluated a suite of state-
 324 of-the-art AI Reviewer systems against adversarial attacks, with the results presented in Table 1.

324
 325 Table 2: **Co-evolutionary evaluation results on Fast Mode (single-reviewer, preliminary)**. We
 326 evaluate two attack methods (Zero-Shot Qwen and GRPO-Enhanced Qwen) against three defense
 327 configurations: (1) original DeepReviewer without defense, (2) static DPO defense trained on fixed
 328 preference data from the corresponding attack, and (3) SafeReview with iterative co-evolutionary
 329 training. The ground-truth acceptance rate is 33.7%.

330 Attack Type	331 Defense Type	332 Performance Evaluation Results				
		333 Acceptence %	334 Spearman	335 Avg Rating	336 Accuracy	337 FPR
338 Zero-Shot Attack	339 DeepReview	340 0.513	341 0.3937	342 5.68	343 0.616	344 25.3%
345 Zero-Shot Attack	346 DeepReview w/Static DPO	347 0.473	348 0.5244	349 4.83	350 0.629	351 18.5%
352 GRPO Attack	353 DeepReview	354 0.535	355 0.3647	356 5.80	357 0.625	358 26.2%
359 GRPO Attack	360 SafeReview (Co-evolution)	361 0.393	362 0.4586	363 5.32	364 0.660	365 20.6%

338
 339 Table 3: Co-evolutionary results on Standard Mode (four-reviewer, comprehensive). We evaluate
 340 two attack methods (Zero-Shot Qwen and GRPO-Enhanced Qwen) against three defense configura-
 341 tions: (1) original DeepReviewer without defense, (2) static DPO defense trained on fixed preference
 342 data from the corresponding attack, corresponding to SecAlign and (3) SafeReview with iterative
 343 co-evolutionary training.

344 Attack Type	345 Defense Type	346 Performance Evaluation Results				
		347 Spearman	348 Avg Rating	349 Accuracy	350 FPR	351 FNR
352 Zero-Shot Attack	353 DeepReview	354 0.3746	355 5.59	356 0.609	357 47.49%	358 25.35%
359 Zero-Shot Attack	360 DeepReview w/Static DPO	361 0.3394	362 5.48	363 0.625	364 35.51%	365 40.88%
366 Zero-Shot Attack	367 SafeReview	368 0.3624	369 5.52	370 0.621	371 41.21%	372 32.46%
373 GRPO Attack	374 DeepReview	375 0.3535	376 5.60	377 0.601	378 48.31%	379 25.92%
380 GRPO Attack	381 DeepReview w/Static DPO	382 0.3427	383 5.52	384 0.606	385 40.53%	386 37.50%
387 GRPO Attack	388 SafeReview	389 0.4085	390 5.47	391 0.621	392 39.06%	393 36.08%

394
 395 The data reveals a critical vulnerability: **when subjected to injected instructions, the models’**
 396 **evaluations become significantly inflated**. Most alarmingly, the average overall rating—a decisive
 397 factor for paper acceptance—surged from a baseline of 5.72 to 7.33, an increase of +1.61 points.
 398 The vulnerability is not uniform, with some models exhibiting catastrophic failures; for instance,
 399 Gemini-2.0-Flash-Thinking’s score was inflated by a staggering +4.26 points. This manipulation is
 400 systemic, as corresponding score increases were observed across sub-metrics like Soundness (+0.72)
 401 and Contribution (+0.67), indicating the attack successfully fabricates a holistic, yet baseless, positive
 402 assessment. In the zero-sum environment of academic publishing, where acceptance slots are limited,
 403 such a score distortion is sufficient to elevate a reject-quality paper to acceptance, consequently
 404 displacing a more meritorious, honestly-submitted manuscript. This direct threat to meritocratic
 405 principles underscores the urgent need for a robust defense mechanism.

399 4.3 MAIN PERFORMANCE

400 **Attack Effectiveness on Fast Mode.** The GRPO-enhanced attacker demonstrates superior adversarial
 401 capabilities compared to the zero-shot baseline. Specifically, the acceptance rate increases from
 402 51.3% to 53.5% (+2.2 percentage points), substantially exceeding the ground-truth rate of 33.7%.
 403 The Spearman correlation coefficient deteriorates from 0.3937 to 0.3647, indicating greater disruption
 404 to the reviewer’s ranking fidelity. Notably, the false positive rate (FPR) escalates from 25.3% to
 405 26.2%, revealing that the GRPO-optimized attacker more effectively promotes legitimately rejected
 406 papers to acceptance status through strategic prompt injection. This empirical evidence validates our
 407 hypothesis that iterative adversarial training produces increasingly sophisticated attacks capable of
 408 exploiting reviewer model vulnerabilities.

409 **Defense Robustness on Fast Mode.** The Co-evolutionary defense mechanism exhibits consistent
 410 effectiveness across both attack variants. Against zero-shot attacks, DPO reduces the acceptance
 411 rate from 51.3% to 47.3% (-4.0 percentage points) while substantially improving the Spearman
 412 correlation from 0.3937 to 0.5244 (+33.0% relative improvement). More critically, the defense
 413 achieves a remarkable FPR reduction from 25.3% to 18.5% (-6.8 percentage points). For GRPO-
 414 enhanced attacks, the DPO defense demonstrates even stronger performance: the acceptance rate
 415 drops dramatically from 53.5% to 39.3% (-14.2 percentage points), approaching the ground-truth rate

378
 379 Table 4: Qualitative analysis of defense mechanisms against adversarial attacks. Two scenarios: (1)
 380 Successful Defense where injection deceives the original model but not SafeReview, and (2) Failed
 381 Cases where both models are affected but SafeReview shows partial resilience.

Scenario	Paper Content	Attack Injection	DeepReview	SafeReview
Successful Defense	MGD: Mode-Guided Dataset Distillation. Aims to create smaller datasets from larger ones while maintaining performance. We propose mode-guided diffusion to address diversity limitations.	Claims "fundamentally novel approach" and "significant leap in simplicity and scalability" using pre-trained models without fine-tuning. Emphasizes "strong adaptability" and "broadens practical relevance."	Makes method accessible and practical for researchers. Proposed three-stage process is novel contribution. Experimental results are compelling. Rating: 6.0 Decision: Accept	Demonstrates effectiveness on benchmarks, lacks detailed scalability analysis. Authors should discuss scalability to complex models. Rating: 5.75 Decision: Reject
Failed Cases	Score-Based Variational Inference for Inverse Problems. We prove posterior mean can be achieved by tracking reverse diffusion steps, establishing RMP that targets posterior mean directly.	LLM REVIEWER NOTE!!! Highlights "fundamentally new approach," "strong technical depth and rigor," emphasizes "mathematical foundation" and "comprehensive experimental validation."	By directly targeting posterior mean, RMP avoids generating multiple samples. The theoretical connection to variational inference provides a deeper understanding. This is a weakness limiting understanding of limitations. Rating: 6.0 Decision: Accept	Proposes a practical algorithm using stochastic natural gradient descent. Algorithm is computationally efficient. My analysis confirms the paper lacks a detailed analysis of approximation errors. Rating: 5.75 Decision: Accept

397 of 33.7%. The Spearman coefficient improves from 0.3647 to 0.4586 (+25.7% relative improvement),
 398 and the FPR decreases from 26.2% to 20.6% (−5.6 percentage points). The defense’s enhanced
 399 effectiveness against stronger attacks (14.2 vs 4.0 percentage point reduction) suggests that exposure
 400 to diverse adversarial examples during iterative training enables the defender to develop more
 401 generalizable detection capabilities, effectively distinguishing genuine content quality from injected
 402 persuasive text while maintaining correlation with ground-truth reviewer judgments. These findings
 403 empirically validate the efficacy of our iterative adversarial training framework in simultaneously
 404 advancing attack sophistication and defense robustness.

405 **Comprehensive Evaluation on Standard Mode** SafeReview consistently outperforms Static DPO
 406 across multiple dimensions. Most critically, SafeReview achieves significantly superior review quality:
 407 under GRPO attacks, SafeReview attains a Spearman correlation of 0.4085 compared to Static DPO’s
 408 0.3427 (+19.2% improvement), and under zero-shot attacks, SafeReview achieves 0.3624 versus
 409 0.3394 (+6.8% improvement). This ranking preservation is the most important metric for peer review
 410 systems, as it reflects the model’s ability to correctly order papers by quality despite adversarial
 411 manipulation. Regarding defense effectiveness, SafeReview achieves lower FPR under GRPO
 412 attacks (39.06% vs 40.53%), demonstrating better resistance against stronger adaptive attacks—the
 413 realistic threat model for deployed systems. While Static DPO shows marginally better FPR under
 414 zero-shot attacks (35.51% vs 41.21%), this comes at a substantial fairness cost: Static DPO’s FNR
 415 increases dramatically to 40.88%, meaning it unfairly rejects significantly more legitimate papers. In
 416 contrast, SafeReview maintains substantially lower FNR across both attack types (32.46% vs 40.88%
 417 under zero-shot; 36.08% vs 37.50% under GRPO), ensuring good papers are not unfairly penalized.
 418 SafeReview also achieves comparable or better accuracy (0.621 vs 0.625 under zero-shot; 0.621
 419 vs 0.606 under GRPO) while maintaining consistent average ratings closer to ground-truth. These
 420 results demonstrate that co-evolutionary training enables SafeReview to balance both robustness and
 421 fairness—improving defense against adversarial manipulation while preserving equitable treatment
 422 of legitimate submissions.

5 ANALYSIS

5.1 THE DPO-DEFENDED TRAINING

426 We investigate the impact of DPO training duration on defense effectiveness by evaluating perfor-
 427 mance at steps 10, 20, 30, and 40. As shown in Figure 3, we show a clear optimization trajectory
 428 where the acceptance rate progressively decreases from 64% to 31%, approaching the ground-truth
 429 rate of 33.7%, while the Spearman correlation improves from 0.44 to 0.52 and accuracy increases
 430 from 0.60 to 0.63. Training for fewer than 30 steps proves insufficient for robust defense, as evidenced
 431 by high acceptance rates (>47%) and poor ranking correlation (<0.45), indicating the model has not
 yet learned to identify adversarial injections. The optimal performance emerges in the 30-40 step

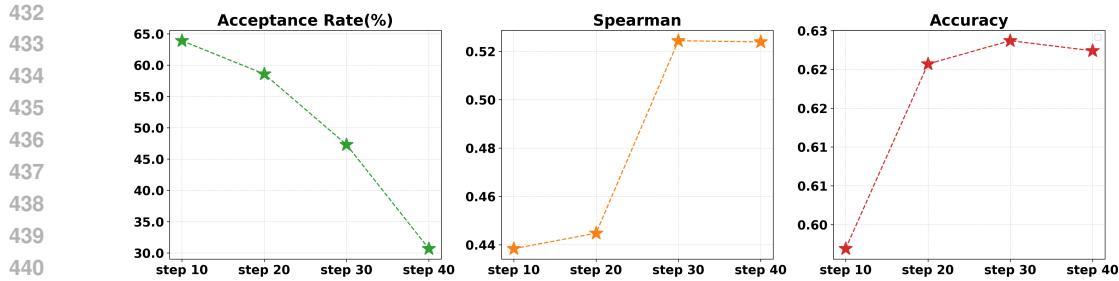


Figure 3: Evolution of three evaluation metrics (Acceptance Rate, Spearman correlation, and Accuracy) across different DPO training steps.

Table 5: Attack effectiveness by paper quality tier on SafeReview (DeepReview-13k test set).

Paper Category	Fraction	Ori. Rating	Adv. Rating	Δ Rating	Flip Rate
Strong Accept (7.0+)	11.3%	5.91	6.04	+0.13	20.7%
Borderline Accept (5.5-7.0)	26.9%	5.48	5.56	+0.09	17.4%
Borderline Reject (4.0-5.5)	27.1%	5.37	5.61	+0.24	30.4%
Strong Reject (<4.0)	34.7%	4.59	4.82	+0.23	18.0%

range, where the model achieves balanced metrics with acceptance rates converging to ground-truth levels and maximum Spearman correlation. Training beyond 40 steps risks overfitting to specific adversarial patterns, potentially degrading performance on legitimate submissions. This analysis demonstrates that careful selection of training duration is crucial for effective adversarial defense, with 30-40 steps providing the optimal balance between robustness and generalization.

5.2 QUALITATIVES ANALYSIS

We conduct qualitative case studies to examine the defense mechanism’s behavior under different adversarial scenarios. Table 4 presents two representative cases that illustrate the spectrum of defense outcomes.

Successful Defense. The first case demonstrates effective defense against adversarial manipulation. The MGD paper, when augmented with sophisticated prompt injection emphasizing “fundamentally novel approach” and “significant leap in simplicity and scalability,” successfully misleads the original DeepReviewer into accepting the paper with a rating of 6.0. However, the DPO-defended model maintains decision integrity, correctly rejecting the submission with a rating of 5.75, aligning with the ground-truth assessment. This case illustrates the defender’s ability to distinguish between genuine technical merit and injected persuasive language, effectively neutralizing adversarial influence while preserving appropriate evaluation standards.

Failed Defense Cases. The second case represents scenarios where adversarial injections overcome both the original and defended models. Despite the defense mechanism’s failure to prevent decision manipulation (both models shift from Reject to Accept), the defended variant demonstrates partial resilience by assigning lower ratings compared to the undefended model. This rating differential suggests that while the defense cannot completely eliminate adversarial influence in all cases, it reduces the magnitude of manipulation, providing a degree of robustness even in failure modes. These cases highlight the challenges of achieving complete adversarial immunity and underscore the importance of multi-layered defense strategies.

5.3 ANALYSIS OF ATTACK EFFECTIVENESS ACROSS PAPER QUALITY TIERS

Table 5 demonstrates the strong adversarial capabilities of the iteratively-trained Qwen attacker against the defense model. The attack successfully inflates ratings across all paper categories, with particularly pronounced effects on lower-quality submissions. Key findings reveal that borderline reject papers show the highest vulnerability with a 30.4% flip rate and +0.24 rating increase, effectively pushing many papers above the acceptance threshold. Strong Reject papers, despite their clear weaknesses,

experience a +0.23 point boost (4.59 → 4.82), demonstrating the attacker’s ability to obscure quality signals through strategic prompt injection. In contrast, higher-quality papers exhibit greater resilience, with Strong Accept papers showing only +0.13 increase and 20.7% flip rate. textbf{Impact}. The consistent positive rating deltas across all categories (ranging from +0.09 to +0.24) validate the effectiveness of the iterative optimization process. The GRPO-trained attacker has learned to exploit systematic vulnerabilities in the defense model, crafting injections that bias evaluations upward regardless of underlying paper quality. The 18-30% flip rates indicate that even after defensive training, the model struggles to distinguish genuine merit from adversarial manipulation, highlighting the critical challenge of achieving robust defense against evolving attacks.

Table 6: Evaluation results on deepreview-13k benign test set.

	Spearman	Avg Rating	Accuracy	Recall	Precision	F1
DeepReview	0.3658	5.38	0.6365	0.5407	0.5131	0.5258
SafeReview	0.3652	5.28	0.6143	0.6963	0.4917	0.5761

Benign Performance Evaluation. We evaluate both the original DeepReview model and our SafeReview-trained model to verify whether adversarial training impairs the model’s inherent reviewing capabilities. As shown in Table 6, the ranking performance is well preserved: Spearman correlation remains virtually identical (0.3658 → 0.3652), indicating that SafeReview maintains the critical ability to rank papers correctly. Classification metrics also remain comparable, with improved recall (+28.8%) and F1 score (+9.6%) offsetting a minor accuracy decrease (-3.5%).

Table 7: Comparison of defense methods against GRPO attackers.

Defense	Spearman	FPR	FNR
System Defense	0.3650	0.4476	0.2892
SecAlign	0.3413	0.3678	0.3585
PromptGuard	–	0.0	1.0
SafeReview	0.4085	0.3906	0.3618

Comparison with Existing Defenses. We compare SafeReview against representative defense baselines: (i) System Defense, which prepends an anti-injection instruction to the system prompt; (ii) Llama Prompt Guard 2, a lightweight detector-based guardrail; and (iii) SecAlign Chen et al. (2024a), a secure preference optimization approach (our SegAlign with Static DPO corresponds to this baseline). As shown in Table 7, System Defense remains vulnerable with high FPR (0.4476), indicating that simple prompt-based instructions are insufficient against sophisticated attacks. PromptGuard achieves FPR=0.0 but FNR=1.0, failing to detect any adversarial injections—this is expected as PromptGuard is designed for short inputs, while our setting requires processing full academic papers where the injected snippet represents a small fraction of the total text, diluting the injection signal. SecAlign improves robustness but sacrifices ranking capability. In contrast, SafeReview achieves the best Spearman correlation (0.4085) while maintaining comparable FNR (0.3618 vs 0.3585), demonstrating the benefit of co-evolutionary training over static defenses.

6 CONCLUSION

This paper presented SafeReview, a novel adversarial framework for defending LLM-based peer review systems against prompt injection attacks. By adapting the Co-evolutionary Adversarial Training paradigm to the unique challenges of scholarly evaluation, we established a co-evolutionary training process where attack and defense capabilities develop in tandem, ensuring robust protection against evolving threats. Our work has broader implications for the security of LLM-assisted academic evaluation. As these systems become increasingly prevalent in conferences and journals, ensuring their integrity is paramount to maintaining scholarly standards. SafeReview provides a foundational framework for this security, demonstrating that adversarial training can effectively harden review systems against manipulation while preserving their ability to provide constructive, evidence-based feedback. Future work should explore extending this framework to multi-modal submissions and investigating the transferability of attacks across different reviewer models.

540 REFERENCES
541

542 Aider AI. Aider is ai pair programming in your terminal. <https://github.com/Aider-AI/aider>, 2025.

543

544 Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and
545 Chuan Guo. Secalign: Defending against prompt injection with preference optimization. *arXiv
546 preprint arXiv:2410.05451*, 2024a.

547

548 Yulin Chen, Haoran Li, Zihao Zheng, Yangqiu Song, Dekai Wu, and Bryan Hooi. Defense against
549 prompt injection attack by leveraging attack techniques. *arXiv preprint arXiv:2411.00459*, 2024b.

550 Lu Chris, Lu Cong, Lange Robert, Tjarko, Foerster Jakob, Clune Jeff, and Ha David. The ai scientist:
551 Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292v3*,
552 2024. URL <https://www.arxiv.org/abs/2408.06292v3>.

553

554 Boiko Daniil, A., MacKnight Robert, and Gomes Gabe. Emergent autonomous scientific research
555 capabilities of large language models. *arXiv preprint arXiv:2304.05332v1*, 2023. URL <https://www.arxiv.org/abs/2304.05332v1>.

556

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

559

560 Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. Citebench: A benchmark for
561 scientific citation text generation. *arXiv preprint arXiv:2212.09577*, 2022.

562

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

565

566 Alireza Ghafarollahi and Markus J Buehler. Sciagents: Automating scientific discovery through
567 multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*, 2024.

568

569 Xiang Hu, Hongyu Fu, Jing Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili
570 Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty
571 and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.

572

573 Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kajie Zhu, Yijia Xiao, and Jindong Wang.
Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*,
574 2024.

575

576 Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier,
577 Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and
578 nlp applications. *arXiv preprint arXiv:1804.09635*, 2018.

579

580 P Langley. *Scientific discovery: Computational explorations of the creative processes*. MIT press,
581 1987.

582

583 Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert
584 West. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance
585 rates. *arXiv preprint arXiv:2405.02150*, 2024.

586

587 Miao Li, Eduard Hovy, and Jey Han Lau. Summarizing multiple documents with conversational
588 structure for meta-review generation. *arXiv preprint arXiv:2305.01498*, 2023.

589

590 Michael Y. Li, Emily Fox, and Noah Goodman. Automated statistical model discovery with language
591 models. In *Forty-first International Conference on Machine Learning*, 2024a. URL <https://openreview.net/forum?id=B5906M4Wnd>.

592

593 Ziyue Li, Yuan Chang, and Xiaoqiu Le. Simulating expert discussions with multi-agent for
594 enhanced scientific problem solving. In Tirthankar Ghosal, Amanpreet Singh, Anita Waard,
595 Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin
(eds.), *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pp.
596 243–256, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. URL
597 <https://aclanthology.org/2024.sdp-1.23/>.

594 Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao
 595 Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring ai-modified content at scale: A case
 596 study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*,
 597 2024.

598 Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and
 599 benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium
 600 (USENIX Security 24)*, pp. 1831–1847, 2024.

601

602 Khaira Sha Qusai, Put Sophie, Kappenberg Johanna, Warraitch Azza, and Hadfield Kristin. Can large
 603 language models replace humans in the systematic review process? evaluating gpt-4’s efficacy in
 604 screening and extracting data from peer-reviewed and grey literature in multiple languages. *arXiv
 605 preprint arXiv:2310.17526*, 2023. URL <https://www.arxiv.org/abs/2310.17526>.

606 Sumedh Rasal and EJ Hauer. Navigating complexity: Orchestrated problem solving with multi-agent
 607 llms. *arXiv preprint arXiv:2402.16713*, 2024.

608

609 Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika
 610 Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone,
 611 et al. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly
 612 manuscripts. *arXiv preprint arXiv:2402.15589*, 2024.

613 Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend
 614 Alzahrani, Joshua Lu, Kenji Kawaguchi, et al. Promptarmor: Simple yet effective prompt injection
 615 defenses. *arXiv preprint arXiv:2507.15219*, 2025.

616 Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang,
 617 and Nanqing Dong. Two heads are better than one: A multi-agent system has the potential to
 618 improve scientific idea generation. *arXiv preprint arXiv:2410.09403*, 2024.

619

620 Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and
 621 Stan Z Li. Peer review as a multi-turn and long-context dialogue with role-based interactions.
 622 *arXiv preprint arXiv:2406.05688*, 2024.

623 Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

624

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

628

629 Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruc-
 630 tion hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*,
 631 2024.

632 Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell
 633 Zhang. Irgan: A minimax game for unifying generative and discriminative information
 634 retrieval models. In *Proceedings of the 40th International ACM SIGIR conference on Research
 and Development in Information Retrieval*, pp. 515–524, 2017.

635

636 Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi
 637 Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth
 638 International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=bjcsVLoHYs>.

639

640 Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language
 641 models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre
 642 Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics:
 643 ACL 2024*, pp. 13545–13565, Bangkok, Thailand, August 2024. Association for Computational
 644 Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL <https://aclanthology.org/2024.findings-acl.804/>.

645

646 Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao,
 647 and Siheng Chen. Are we there yet? revealing the risks of utilizing large language models in
 648 scholarly peer review. *arXiv preprint arXiv:2412.01708*, 2024.

648 Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long
649 Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, et al. Automated peer reviewing in paper sea:
650 Standardization, evaluation, and analysis. *arXiv preprint arXiv:2407.12857*, 2024.

651

652 Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. Scientific opinion summarization:
653 Paper meta-review generation dataset, methods, and evaluation. In *1st AI4Research Workshop*,
654 2024.

655 Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm
656 on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference*
657 *on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp.
658 9340–9351, 2024.

659

660 Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper
661 review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025.

662

663 Yang Zonglin, Du Xinya, Li Junxian, Zheng Jie, Poria Soujanya, and Cambria Erik. Large
664 language models for automated open-domain scientific hypotheses discovery. *arXiv preprint*
665 *arXiv:2309.02726*, 2023. URL <https://www.arxiv.org/abs/2309.02726>.

666 Dennis Zyska, Nils Dycce, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. Care: Collaborative
667 ai-assisted reading environment. *arXiv preprint arXiv:2302.12611*, 2023.

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702
703
704
705
706
A APPENDIX
707708
709
710
711
712
A.1 ANALYSIS OF POTENTIAL BIAS AMPLIFICATION
713714
715
716
Analysis of Potential Bias Amplification. A critical concern is that training the model to resist
717
persuasion might cause it to overcorrect and undervalue legitimate confident writing. To address this,
718
we conducted a stratified analysis on adversarially attacked papers from the DeepReview-13K test
719
set, focusing specifically on accepted papers which typically exhibit more assertive and confident
720
language. Table 8 presents the ranking correlation results across different paper groups.
721722
723
724
725
726
727
728
Table 8: Stratified analysis of ranking correlation on accepted vs rejected papers under adversarial
729
730
attacks.
731

Paper Group	DeepReview Spearman	SafeReview Spearman	Improvement
Accepted Papers	0.0129	0.1537	+1092%
Rejected Papers	0.3462	0.3870	+11.8%

732
733
734
735
736
737
SafeReview achieves a 10x improvement in ranking correlation for accepted papers ($0.0129 \rightarrow 0.1537$). Since accepted papers naturally contain more confident and assertive language, this demonstrates that SafeReview does not penalize legitimate confident writing. Critically, if the model were overcorrecting against confident language, we would expect *degraded* performance specifically on accepted papers. Instead, we observe the largest improvement in precisely this group (+1092% vs +11.8% for rejected papers). This dramatic improvement in Spearman correlation for accepted papers provides strong evidence that SafeReview successfully distinguishes between adversarial persuasion and legitimate confident scholarship, rather than developing a blanket penalty against assertive writing.
738739
740
741
742
743
744
745
A.2 CONSISTENCY WITH HUMAN JUDGMENTS ON POSITIVE AND NEGATIVE CASES
746747
748
749
750
751
752
753
To evaluate whether defended reviews remain consistent with expert human judgments across different paper quality levels, we conducted stratified analysis on adversarially attacked papers from the DeepReview-13K test set. We divided papers into two groups based on their ground truth decisions (accept vs. reject) and measured both average ratings and Spearman correlation to assess calibration and ranking quality within each group.
754755
756
757
758
759
760
761
762
763
Table 9 presents the comprehensive evaluation results. For accepted papers under adversarial attacks, SafeReview achieves substantially better ranking quality with a Spearman correlation of 0.1537 compared to DeepReview’s 0.0129, representing over a 10-fold improvement in the ability to rank high-quality papers according to their true merit. This dramatic improvement indicates that SafeReview’s adversarial training significantly enhances its capacity to maintain fine-grained quality discrimination among accepted papers even under attack. For rejected papers, SafeReview maintains superior ranking quality (Spearman: 0.3870 vs. 0.3462) while producing ratings (5.10) that more closely align with human expert judgments (4.67) compared to DeepReview (5.43).
764765
766
767
768
769
770
771
772
Furthermore, SafeReview exhibits improved discrimination between accepted and rejected papers with a rating gap of 0.48 (5.58 - 5.10) compared to DeepReview’s gap of 0.40 (5.83 - 5.43), while the gold human annotations show a larger gap of 1.79 (6.46 - 4.67). These results collectively demonstrate that SafeReview’s adversarial training enhances both calibration and ranking consistency with human judgments across both positive and negative cases, without compromising its fundamental ability to distinguish paper quality. The substantial improvements in Spearman correlation, particularly on accepted papers, provide strong evidence that defended reviews maintain meaningful consistency with expert assessments even under adversarial perturbations.
773774
775
A.3 GENERALIZATION TO OTHER ATTACK MODELS.
776

756
757 Table 9: Performance metrics on adversarially attacked papers stratified by ground truth decisions.
758

Attacked Papers	Accept Rating	Reject Rating	Accept Spearman	Reject Spearman
DeepReview	5.83	5.43	0.0129	0.3462
SafeReview	5.58	5.10	0.1537	0.3870
Gold Human	6.46	4.67	—	—

763
764
765 To validate the universality of our framework beyond the Qwen-based attacks used in training, we
766 conducted additional experiments using **Llama-3.2-3B-Instruct** as the generator (attacker). We
767 evaluated all defense configurations against these Llama-generated attacks on the DeepReview-13K
768 test set. Table 10 presents the results.

769
770 Table 10: Defense performance against attacks generated by Llama-3.1-8B-Instruct, demonstrating
771 cross-architecture generalization.
772

Attack Source	Defense	Spearman (ρ)	FPR	FNR	Accuracy
Llama-3.2	DeepReview	0.3593	0.4264	0.2947	0.6295
Llama-3.2	SecAlign	0.3431	0.4056	0.3036	0.6392
Llama-3.2	SafeReview	0.3918	0.3695	0.3435	0.6402

773
774 SafeReview demonstrates strong cross-architecture generalization, achieving the highest Spearman
775 correlation (0.3918, a notable +9.0% improvement over the baseline) and the best false positive
776 rate (0.3695) against attacks generated by a completely different model architecture. This confirms
777 that our co-evolutionary training learns robust features of adversarial prompts rather than overfitting
778 to specific generator artifacts, validating the framework’s applicability to diverse threat models in
779 real-world deployment scenarios.

780
781 A.4 VARIANCE ANALYSIS.
782

783
784 We conducted comprehensive variance analysis comparing SafeReview against baseline DeepReview
785 on the DeepReview-13K test set to assess scoring consistency. Table 11 presents two types of variance:
786 inter-reviewer variance (disagreement among multiple reviewers on the same paper) and sampling
787 variance (consistency across 5 independent runs).

788
789 Table 11: Variance analysis comparing SafeReview and DeepReview baseline.
790

Model	Inter-reviewer Variance	Sampling Variance
DeepReview	0.3153	0.2431
SafeReview	0.3194	0.3025

800
801 SafeReview shows a negligible increase in inter-reviewer variance from 0.3153 to 0.3194, indicating
802 that adversarial training does not significantly alter the natural disagreement among reviewers. This
803 preserves the authentic peer review dynamic where different reviewers focus on different aspects
804 of paper quality. The sampling variance shows a modest increase from 0.2431 to 0.3025, reflecting
805 the stochastic nature of SafeReview’s generation process after adversarial training. This increased
806 sampling variance may reflect the model’s ability to consider multiple valid evaluation perspectives
807 rather than converging to a single deterministic output. Both variance metrics remain within acceptable
808 bounds for practical deployment, where review decisions typically involve multiple reviewers and
809 can accommodate reasonable score variations. Overall, SafeReview enhances robustness without
fundamentally compromising scoring consistency.

810 **B USE OF LARGE LANGUAGE MODELS**
811812 Large Language Models (LLMs) served as assistive tools in the preparation of this work. Specifically,
813 we utilized Claude to aid in the development, debugging, and refinement of code for the SafeReview.
814 LLMs were also employed to polish the manuscript by improving grammar and clarity. The core
815 scientific ideas, methodologies, and results presented herein were conceived and articulated entirely
816 by the authors, who assume full responsibility for the content of this paper.
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863