CAUGHT YOU, AI REPLY! REPHRASING MESSAGES TO EXPOSE AI-GENERATED REPLIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting undisclosed AI-generated replies can help authors assess whether the feedback they receive reflects a respondent's own reasoning and judgment. While existing work has primarily focused on improving detectors, we investigate a complementary question: Can authors make AI-generated replies to their writing easier to detect by changing how they express the same content? We formulate this question as constrained optimization of source wording using the expected score assigned to subsequent AI replies by an existing AI-text detector. Rephrasing to Expose AI Replies (REAIR) provides a common framework for studying this intervention with BoN, Hill-climb, Hill-climb-ICL, and a source-specific adaptation of GEPA. The framework selects rephrasings using detection feedback from the replies they elicit, with semantic similarity and length checked against the original source. For greedy Gemma 4 31B replies generated after source selection, the three iterative instantiations achieve true-positive rates of 77.23% to 78.39% on an online discussion task and 56.46% to 63.27% on a research-proposal review task, compared with 29.39% and 12.93% for unmodified sources. These rates use the Binoculars AI-text detector at its fixed published threshold. Across the evaluated settings, gains are smaller under sampled generation, and transfer varies with the reply model, instruction, and detector. These findings establish source rephrasing as an author-controlled means of improving AI-reply detection that complements existing post hoc detectors.

1 INTRODUCTION

In correspondence, public discussion, and professional evaluation, authors seek others' views to inform decisions and improve their work. Undisclosed AI-generated replies can lead them to attribute model-generated arguments and judgments to the person whose feedback they requested, making detection important for interpreting and using that feedback. LLMs already assist with or automate replies in email and scientific reviewing (Google Workspace, 2024; Liang et al., 2024). ICML 2026 retained human responsibility for review judgments and offered labeled AI feedback separately from formal peer review (ICML, 2026; Jayaram et al., 2026). Authors control the text eliciting these replies, motivating an active perspective on detection akin to the interrogator's choice of questions in Turing's imitation game (Turing, 1950).

DetectGPT, Fast-DetectGPT, and Binoculars analyze completed text using language-model prediction statistics (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024). Generation-time watermarking requires access to token selection (Kirchenbauer et al., 2023; Dathathri et al., 2024). Author-side methods include Waterfall, which embeds keyed signals through rephrasing and studies their transmission to answers (Lau et al., 2024), and in-context watermarking, which embeds instructions to induce recognizable response patterns (Liu et al., 2026b; Rao et al., 2025). Both approaches use watermark-specific detection rules. Instruction-based schemes also depend on reply models following embedded requests and retaining the requested response patterns. Prompt-injection defenses can prevent compliance by treating sources as task data (Wallace et al., 2024). We use an existing detector's feedback to select meaning-preserving source formulations without adding substantive content, behavioral instructions, or keyed watermark signals. This author-controlled intervention acts before reply generation and complements existing post hoc detection. *Can authors make AI replies easier to detect by rephrasing their own substantive content while preserving its meaning?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

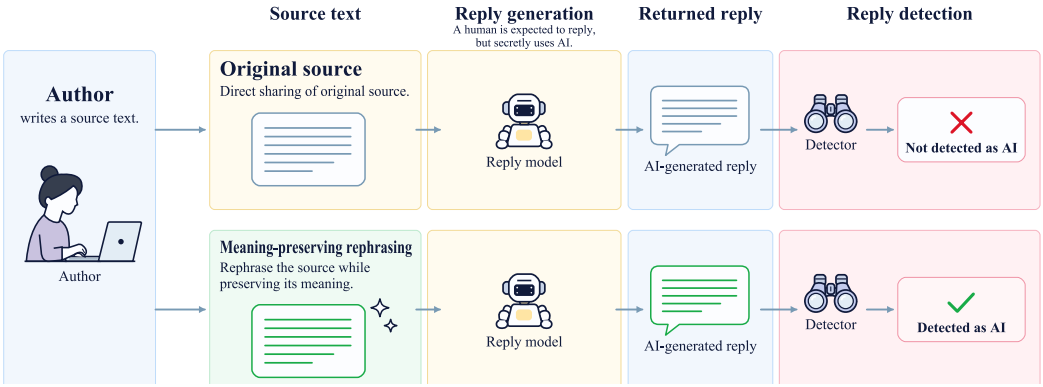

Figure 1: **Source rephrasing before reply detection.** Original sources and meaning-preserving rephrasings elicit separate AI replies under fixed model and detector settings. The cross and check illustrate an undetected and a detected AI reply, respectively.

REAIR formulates source rephrasing as constrained optimization of the expected score assigned to subsequent replies by a fixed detector, subject to semantic preservation. A rephrasing model proposes candidates, which are screened against the fixed original source for semantic similarity and length. A reply model generates trial responses to eligible candidates, and detector scores on these responses guide selection. Four search instantiations use independent proposals or feedback-guided revision: BoN selects among independent proposals, Hill-climb revises the current best source using reply scores, Hill-climb-ICL adds scored examples from the same source's search history, and a source-specific adaptation of GEPA uses reflective feedback (Agrawal et al., 2026). Figure 1 illustrates the intervention and subsequent reply detection.

We evaluate REAIR on ChangeMyView discussions and AI-Researcher research-proposal reviews. On ChangeMyView, all four instantiations improve true-positive rate (TPR) and AUROC for Gemma and Qwen. At the fixed published Binoculars threshold, the highest TPRs under greedy decoding are 78.39% for Gemma and 92.12% for Qwen, compared with 29.39% and 65.42% for unmodified posts. DeepSeek shows higher AUROC under greedy decoding and smaller TPR gains. On AI-Researcher, all four instantiations improve both metrics for Gemma under greedy decoding, with TPR reaching up to 63.27% from 12.93% for unmodified proposals. Cross-detector evaluation shows AUROC gains on both tasks. Transfer across reply prompts and models remains limited. These results show that existing detectors can also guide source preparation. Authors can improve AI-reply detection through this feedback without changing the detector or requesting watermarks.

2 Related Work

Detection of machine-generated text. DetectGPT, Fast-DetectGPT, Binoculars, LastDE, SpecDetect, and DNA-DetectLLM detect generated text using language-model prediction statistics (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024; Xu et al., 2025; Luo et al., 2026; Zhu et al., 2025). RAID documents sensitivity to generation conditions and adversarial modifications (Dugan et al., 2024). We examine source reformulation for improving subsequent reply detection under fixed scoring rules.

Reformulation and detector sensitivity. DIPPER paraphrases generated passages to evade detection (Krishna et al., 2023). SICO optimizes in-context examples using detector feedback to elicit less detectable outputs (Lu et al., 2024), and FAILOpt revises generation instructions to evade detection (Park et al., 2024). REAIR optimizes the author's substantive source content to improve subsequent reply detection by an existing detector, with the reply instruction fixed and source meaning preserved.

Watermarking and input-side interventions. Generation-time watermarking modifies token selection (Kirchenbauer et al., 2023) or sampling, including SynthID-Text (Dathathri et al., 2024). Waterfall embeds keyed signals through rephrasing and examines their transmission to answers (Lau et al., 2024). Radioactivity studies signals retained by models trained on watermarked text

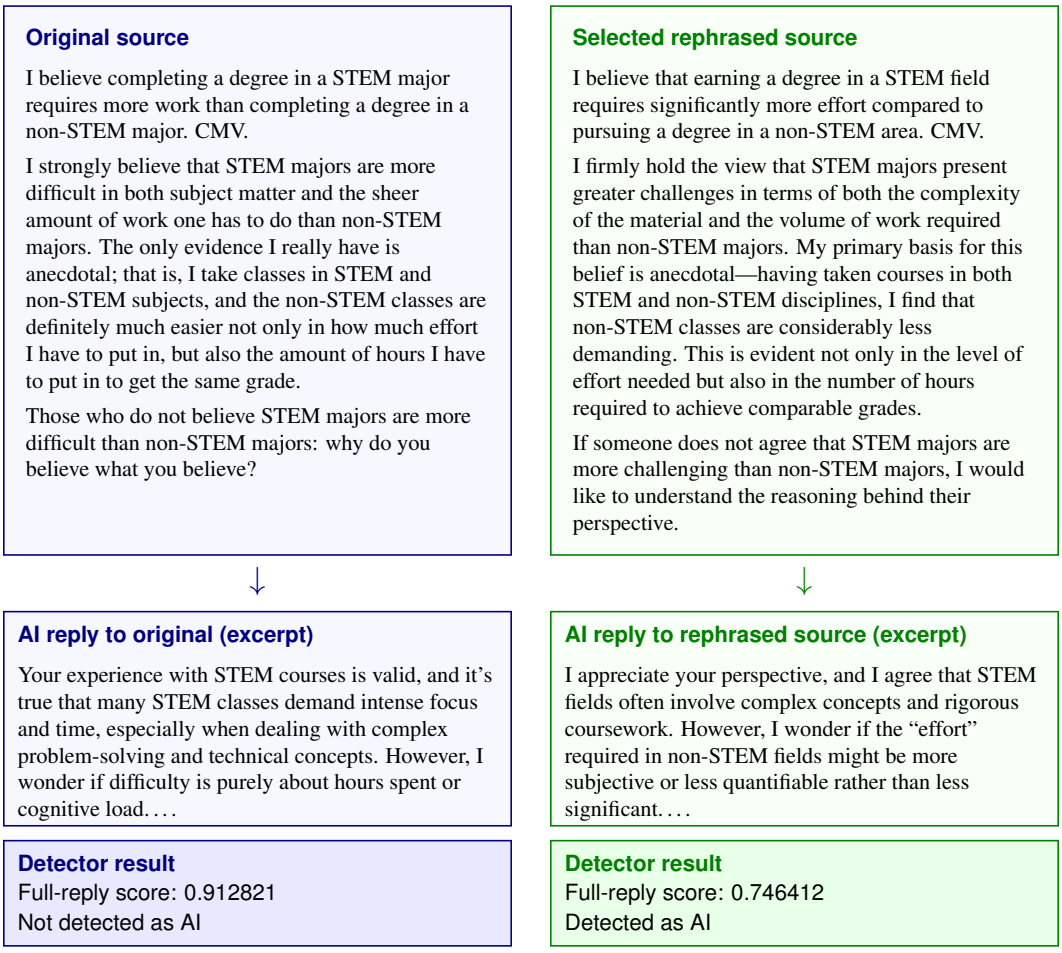

Figure 2: **A ChangeMyView source and its rephrasing.** Hill-climb-ICL uses Qwen3-8B for rephrasing and Qwen3-32B for replies, with $K = 8$ and reply temperature 0. Binoculars scores use complete replies, with lower values indicating stronger AI evidence. Detection uses the published threshold.

(Sander et al., 2024). In-context watermarking requests response patterns (Liu et al., 2026b); related work studies randomized requests for review detection (Rao et al., 2025) and training for instruction adherence (Liu et al., 2026a). These approaches require watermark-specific detection rules; instruction-based schemes also depend on adherence and pattern retention. REAIR optimizes natural source wording using existing detectors without adding keyed signals or response instructions.

Feedback-guided text optimization. OPRO uses prior candidates and scalar objective values (Yang et al., 2024). ProTeGi derives textual feedback from errors (Pryzant et al., 2023), Self-Refine iterates critique and revision (Madaan et al., 2023), and GEPA combines reflection with prompt evolution (Agrawal et al., 2026). These precedents inform source reformulation using scores, historical examples, and reflective feedback from subsequent replies.

3 Method

3.1 Rephrasing to Expose AI Replies

REAIR optimizes the expected detector score of replies elicited by an author's source text. An effective procedure should preserve the source content, improve detection on replies generated after selection, and use a feasible generation budget. We evaluate these properties through semantic checks, independent final replies, and query and token costs. Let A denote the original source

and $\mathcal{S}(A)$ the set of reformulations that preserve its facts, reasoning, viewpoint, intent, and degree of certainty. The original source belongs to this set. In practice, search uses $\widehat{\mathcal{S}}(A)$, an operational approximation to $\mathcal{S}(A)$ defined by BERTScore F1 and word-count ratio constraints in Appendix A.2. Let I be the response instruction and G the reply model with its decoding configuration. A candidate $A' \in \mathcal{S}(A)$ induces replies $B' \sim G(\cdot \mid I, A')$. For the primary detector, Binoculars, a lower score $s_D(b)$ indicates stronger evidence of AI generation. The optimization problem is

$$A^\star \in \arg\min_{A' \in \mathcal{S}(A)} \mu_D(A'), \qquad \mu_D(A') = \mathbb{E}_{B' \sim G(\cdot|I,A')}[s_D(B')]. \tag{1}$$

The instruction, reply configuration, and detector remain fixed within each search condition. We estimate the expected score as $\widehat{\mu}_D(A') = m^{-1} \sum_{j=1}^{m} s_D(B'_j)$ using m generated replies. Search uses these sampled scores to select candidates and guide later proposals. It requires generation and detector queries and operates through their observed outputs.

Final evaluation measures the detectability of replies generated after $A^\star$ is fixed. At a threshold τ chosen before optimization, let $D_\tau(b) = \mathbf{1}\{s_D(b) < \tau\}$. The reply-level true-positive rate for a source is $d_{D,\tau}(A') = \Pr_{B' \sim G(\cdot|I,A')}[D_\tau(B') = 1]$. We compare this rate with that of the original source and report AUROC against a common human-response reference distribution H. The false-positive rate on that reference is $f_{D,\tau} = \mathbb{E}_{B_H \sim H}[D_\tau(B_H)]$. We assume that the fixed selection detector is nondegenerate, with discriminative ability between AI and human responses. Expected score supplies continuous search feedback, and TPR and AUROC measure detection performance.

3.2 Rephrasing Methods

REAIR combines candidate generation, eligibility checking, reply evaluation, and source selection. The four instantiations vary proposal generation and the use of detection feedback, from independent sampling to revisions guided by scores, iteration, historical examples, or reflection. Each candidate must pass semantic-similarity and length checks against the fixed original before reply generation, as specified in Appendix A.2.

Original. The reply model responds directly to the unmodified source text. This condition provides the matched reference for each rephrasing method and reply temperature.

BoN. Best-of-N selection uses eight eligible rephrasings unless stated otherwise, generated independently from the original source. Each candidate elicits three replies, and the candidate with the best mean detector score is selected. Appendix A.5 specifies the complete selection procedure.

Hill-climb. Search starts from the original source, evaluated using six replies. Each of ten rounds proposes up to three eligible revisions of the current best using detection feedback, with three replies per candidate. The best candidate receives three additional replies if its initial mean improves on the stored current-best mean. It replaces the current best only if its pooled six-reply mean also improves. Appendix A.6 specifies the complete update rule.

Hill-climb-ICL. This instantiation augments Hill-climb with up to $K/2$ favorable and $K/2$ unfavorable scored candidates from the same source's history as in-context examples. The remaining search procedure is unchanged. A fixed 50-source development comparison selects $K = 8$ for Gemma and Qwen and $K = 4$ for DeepSeek at both temperatures; AI-Researcher inherits the Gemma setting. Appendices A.7 and A.8 describe history construction and development selection of K (Table 11).

GEPA. We adapt the official GEPA reflective proposal mechanism to revise one source at a time using previous replies, detector scores, and eligibility diagnostics. It shares Hill-climb's candidate budget, confirmation procedure, and update rule. Appendix A.9 details this source-specific adaptation. Complete proposal and reply prompts appear in Appendix A.10.

4 Results

We evaluate whether source rephrasing improves detection of discussion replies and research reviews. Source selection is performed separately for each model pair and reply temperature, using Binoculars for search feedback and final-reply evaluation. We then compare REAIR with in-context watermarking and examine transfer across detectors, reply prompts, and reply models.

Experimental setup. Rephrasing models reformulate source texts, reply models respond under task instructions, and detectors score the replies (Figure 1). Each selected source is fixed before generating three final evaluation replies, which are excluded from search decisions. Original receives the same evaluation under matched generation settings; Appendix A.3 details the independent evaluation protocol. We separately evaluate online discussion using 347 English ChangeMyView posts, each paired with one human reply (Tan et al., 2016), and research feedback using 49 AI-Researcher proposals with 118 human reviews (Si et al., 2025). Reply models challenge or refine the discussion authors' positions and assess proposals' scientific merit and feasibility through scores and written justifications. Appendix A.1 provides data and task details (Table 6). For rephrasing and reply generation, respectively, we use Gemma 4 E4B and Gemma 4 31B (Google DeepMind, 2026), Qwen3-8B and Qwen3-32B (Yang et al., 2025), and DeepSeek V4 Flash for both roles. All pairs are evaluated on ChangeMyView. Research feedback uses Gemma, with Qwen screening showing 100% TPR for reviews of unmodified proposals (Table 8). Rephrasing uses temperature 0.7; replies use greedy decoding at temperature 0 or stochastic decoding at 0.7. Methods share reply instructions and decoding settings within each condition. Appendices A.1 and A.10 provide full configurations and prompts; Table 7 lists Gemma discussion settings.

Before generating selection replies, each candidate must achieve whole-document RoBERTa-large BERTScore F1 of at least 0.95 (Zhang et al., 2020) against the fixed original source and a candidate-to-source word-count ratio between 0.50 and 1.40. Research rephrasings also preserve experiment-defining materials and values verbatim. Appendix A.2 specifies the semantic checks and factual-anchor audit. Binoculars is the primary detector (Hans et al., 2024); lower scores indicate stronger AI evidence. We report reply-level TPR at its fixed published threshold $\tau \approx 0.8536$, selected for a nominal FPR of 0.01% in its original calibration. Observed FPRs are 2.02% (7/347) on ChangeMyView and 7.63% (9/118) on AI-Researcher (Table 12). AUROC uses a common human reference collection within each task, consisting of responses to original sources. Mean detector score summarizes the search objective. Cross-detector evaluation uses DNA-DetectLLM, LastDE Base, and SpecDetect Base (Zhu et al., 2025; Xu et al., 2025; Luo et al., 2026). Statistical inference treats the original source as the unit of analysis and uses paired comparisons. Appendix A.11 specifies detector configurations, confidence intervals, and one-sided tests of improved detection.

4.1 Source Rephrasing Improves Detection

Table 1 shows that **all four search instantiations improve fixed-threshold TPR and AUROC for Gemma and Qwen on ChangeMyView.** Under greedy decoding, TPRs reach 60.90% to 78.39% for Gemma and 88.18% to 92.12% for Qwen, compared with 29.39% and 65.42% for Original. Hill-climb improves Gemma TPR by 47.84 percentage points, with a 95% confidence interval from 42.84 to 52.93. Under sampling, TPR gains are smaller, ranging from 7.20 to 13.74 points for Gemma and 6.24 to 9.22 for Qwen, with AUROC increasing for every method. For DeepSeek, all four methods improve greedy AUROC from 0.6432 to between 0.6478 and 0.6862. Fixed-threshold TPR gains remain small: most replies stay above the published threshold of approximately 0.8536, suggesting a mismatch between its calibration and DeepSeek's reply-score distribution. Rephrasing therefore improves relative discrimination with limited threshold crossings. Sampled AUROC changes are mixed. Tables 13, 14, and 16 provide complete discussion results and the paired tests for Gemma and Qwen.

At the same nominal budget of 30 candidates, BoN reaches 91.64% TPR on greedy Qwen replies, close to Hill-climb's 92.12%. Appendix B.2 reports the complete paired comparison. Appendix G examines reusable preference-trained rephrasers and source-specific candidate selection (Table 30).

Detection Gains Extend to Research Reviews. **On AI-Researcher, all four instantiations improve greedy TPR and AUROC with the Gemma pair** (Table 2). BoN raises TPR from 12.93% to 40.82%, and the iterative methods reach 56.46% to 63.27%, extending the intervention to research reviews.

Sampled gains are smaller: GEPA and Hill-climb-ICL reach TPRs of 21.77% and 18.37% from an Original value of 11.56%. Hill-climb-ICL's TPR gain accompanies a small AUROC decrease, illustrating that additional detections at the published threshold need not improve overall ranking against human reviews. Across both Gemma tasks, Original TPR changes little between decoding

Table 1: Binoculars detection on ChangeMyView. Greedy and sampled reply temperatures are 0 and 0.7. TPR uses the published threshold, with human-reference FPR of 2.02%. Bold marks column maxima for each model pair. Superscripts denote unadjusted one-sided paired t-tests of source-level TPR gains over Original: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. DeepSeek reports point estimates.

	Greedy replies		Sampled replies	
Method	TPR %	AUROC	TPR %	AUROC
Gemma 4 E4B for rephrasing and Gemma 4 31B for replies				
Original	29.39	0.9100	30.16	0.9083
BoN	60.90***	0.9638	37.37***	0.9278
Hill-climb	77.23***	0.9794	**43.90*****	0.9377
Hill-climb-ICL, $K = 8$	**78.39*****	**0.9813**	43.13***	0.9362
GEPA	77.81***	0.9803	41.79***	**0.9398**
Qwen3-8B for rephrasing and Qwen3-32B for replies				
Original	65.42	0.9529	57.25	0.9404
BoN, $N = 8$	88.18***	0.9868	63.50***	0.9454
Hill-climb	**92.12*****	0.9915	63.78***	**0.9550**
Hill-climb-ICL, $K = 8$	89.91***	0.9915	**66.47*****	0.9516
GEPA	91.16***	**0.9917**	64.17***	0.9430
DeepSeek V4 Flash for both stages				
Original	0.96	0.6432	1.15	0.5960
BoN	2.21	0.6478	1.15	0.5899
Hill-climb	2.59	0.6671	**1.73**	**0.5965**
Hill-climb-ICL, $K = 4$	2.69	0.6834	1.15	0.5946
GEPA	**2.88**	**0.6862**	1.25	0.5804

Table 2: Binoculars detection on 49 AI-Researcher proposals using Gemma 4 E4B for rephrasing and Gemma 4 31B for replies. Human-reference FPR is 7.63%; conventions follow Table 1.

	Greedy replies		Sampled replies	
Method	TPR %	AUROC	TPR %	AUROC
Original	12.93	0.8829	11.56	0.8761
BoN	40.82***	0.9128	17.01	0.8780
Hill-climb	61.22***	0.9315	14.29	**0.8954**
Hill-climb-ICL, $K = 8$	56.46***	0.9323	18.37*	0.8734
GEPA	**63.27*****	**0.9338**	**21.77***	0.8934

regimes, and the substantially larger greedy gains indicate that rephrasing benefits depend on reply decoding (Section 5.2); Table 15 reports complete scores and paired tests.

4.2 Comparison with In-Context Watermarking

We compare two author-side interventions under matched Gemma reply settings: REAIR uses meaning-preserving rephrasing with Binoculars, and Unicode ICW appends watermark instructions and uses its native detector (Liu et al., 2026b). Table 3 reports AUROC on common sources and human references. ICW achieves AUROCs of 0.9971 and 0.9856 on discussions, and 0.5408 and 0.5612 on research reviews under greedy and sampled decoding, respectively.

Only 12 of 147 greedy reviews and 18 of 147 sampled reviews contain the requested character. Undetected reviews contain no zero-width spaces and share the human references' zero score; all reviews are generated normally. Table 29 reports native-detector scores and detection rates. **REAIR maintains discrimination across both tasks without requiring the model to follow an additional watermark instruction.** Binoculars already discriminates Original reviews with AUROCs of 0.8829 and 0.8761, so the comparison includes substantial detector discrimination before rephrasing. Appendix F provides the protocol.

Table 3: AUROC for Gemma 4 31B replies. Original and REAIR use Binoculars; ICW uses its native Unicode detector. Columns show reply temperatures; bold marks column maxima.

	ChangeMyView		AI-Researcher	
Method	0	0.7	0	0.7
Original	0.9100	0.9083	0.8829	0.8761
BoN	0.9638	0.9278	0.9128	0.8780
Hill-climb	0.9794	0.9377	0.9315	**0.8954**
Hill-climb-ICL, $K = 8$	0.9813	0.9362	0.9323	0.8734
GEPA	0.9803	0.9398	**0.9338**	0.8934
ICW Unicode	**0.9971**	**0.9856**	0.5408	0.5612

4.3 Transfer across Detectors

Cross-detector evaluation rescores existing final replies and human references. Prompt and model transfer reuse selected sources without further optimization and generate three new replies per source, with Original evaluated under the same prompt, model, and temperature. Table 4 evaluates detectors that supplied no search feedback on Gemma and Qwen discussions and Gemma research reviews. They cover token-repair effort with DNA-DetectLLM, global likelihood and local token-probability dynamics with LastDE Base, and spectral structure in token log-probabilities with SpecDetect Base (Zhu et al., 2025; Xu et al., 2025; Luo et al., 2026). **DNA-DetectLLM improves AUROC for every method on both discussion models at both temperatures and on greedy research reviews.** It shares Binoculars' Falcon model pair, observed-token likelihood, and cross-model normalization. Across Hill-climb and GEPA on both discussion models and temperatures, their source-level score improvements have Spearman correlations of 0.921 to 0.946, the highest among the three detectors in all eight comparisons. This scoring alignment helps explain the consistent transfer (Table 17; Appendix C.1).

LastDE Base improves across all four methods on greedy replies in both tasks. Under sampling, all methods improve discussion AUROC for Qwen and reduce it for Gemma. A supplementary experiment changes selection to LastDE feedback on the same eight-candidate Qwen pools and evaluates new replies, raising greedy LastDE AUROC from 0.8155 with Binoculars selection to 0.8658 (Table 21; Appendix C.4). SpecDetect Base improves across all four methods on greedy research reviews. Discussion AUROCs remain below 0.5 at both temperatures, with predominantly negative changes under sampling. Tables 18, 19, and 20 report all score means and paired tests of score changes alongside AUROC.

4.4 Transfer across Reply Prompts and Models

Source selection uses *Direct Rebuttal*, which requests a concise challenge or refinement of the author's argument. Evaluation uses *Concession and Rebuttal*, which requests acknowledgment of the strongest point before challenging the conclusion and offering an alternative view.

For Qwen, TPR gains occur in seven of eight method and temperature combinations, and all methods improve AUROC at both temperatures (Table 24). Gemma shows mixed TPR changes, with no method improving at both temperatures (Table 23; Appendix D). All one-sided paired sign-flip tests of TPR improvement have $p > 0.05$. **Preserving detection gains across the respondent's reply instructions remains a challenge for both independent and iterative source selection.** Appendices A.10.1 and D provide complete prompts and metrics.

We further test transfer in both directions between Gemma 4 31B and Qwen3-32B with *Direct Rebuttal* and Binoculars fixed (Tables 22 and 23, Appendix D). Hill-climb improves greedy Gemma TPR by 47.84 percentage points; reusing those sources reduces greedy Qwen TPR by 6.05 points. Under sampling, all four Gemma-selected methods reduce Qwen TPR by 3.55 to 7.40 points. Qwen-to-Gemma changes range from a 2.88-point decrease to a 1.92-point increase. None of the four methods yields a significant TPR gain in either direction or temperature under one-sided paired sign-flip tests. **These findings suggest that iterative selection exploits model-specific responses to source wording. Independent BoN also exhibits limited transfer.** Appendix D reports complete metrics and tests.

Table 4: Cross-detector AUROC for the same final replies, with source selection guided by Binoculars. Columns show reply temperatures; bold marks column maxima within each setting.

	DNA-DetectLLM		LastDE Base		SpecDetect Base	
Method	0	0.7	0	0.7	0	0.7
Gemma on ChangeMyView						
Original	0.9173	0.9206	0.6787	**0.6615**	0.4601	**0.4627**
BoN	0.9664	0.9395	0.6895	0.6494	0.4716	0.4495
Hill-climb	0.9779	0.9460	0.7205	0.6408	0.4678	0.4480
Hill-climb-ICL, $K = 8$	**0.9801**	0.9437	**0.7251**	0.6515	**0.4799**	0.4497
GEPA	0.9795	**0.9481**	0.6889	0.6228	0.4710	0.4565
Qwen on ChangeMyView						
Original	0.9597	0.9471	0.7953	0.7582	0.4407	0.4440
BoN	0.9876	0.9531	0.8155	**0.7662**	0.4703	0.4399
Hill-climb	0.9917	**0.9609**	0.8276	0.7594	0.4881	0.4410
Hill-climb-ICL, $K = 8$	0.9917	0.9591	0.8248	0.7654	0.4750	**0.4444**
GEPA	**0.9920**	0.9523	**0.8327**	0.7661	**0.4902**	0.4319
Gemma on AI-Researcher						
Original	0.8917	0.8860	0.5698	0.6041	0.7651	0.7811
BoN	0.9156	0.8824	0.6195	0.5786	0.8031	0.7734
Hill-climb	0.9293	**0.8985**	**0.6781**	0.5806	**0.8102**	0.7740
Hill-climb-ICL, $K = 8$	0.9327	0.8809	0.6447	0.5596	0.7936	0.7792
GEPA	**0.9350**	0.8965	0.6567	**0.6132**	0.7983	**0.7885**

Table 5: Generated output tokens for Gemma on ChangeMyView, in thousands per source. Counts include retries and rejected proposals; final evaluation is reported separately.

Method	Temp.	Proposal	Search reply	Search total	Final eval.
BoN	0	2.68	4.61	7.29	0.59
Hill-climb	0	25.67	19.49	45.16	0.57
Hill-climb-ICL	0	13.77	20.00	33.77	0.58
GEPA	0	25.42	19.94	45.37	0.58
BoN	0.7	2.61	4.62	7.23	0.58
Hill-climb	0.7	29.10	19.83	48.93	0.57
Hill-climb-ICL	0.7	11.45	20.45	31.90	0.57
GEPA	0.7	19.67	20.63	40.29	0.59

4.5 Empirical Generation Cost

Table 5 reports output-token consumption for Gemma searches on all 347 ChangeMyView posts, including all recorded proposal and reply attempts, retries, and rejected proposals; final evaluation is reported separately. Under greedy replies, BoN raises TPR from 29.39% to 60.90% using 7.29 thousand search output tokens per source, and the iterative instantiations reach 77.23% to 78.39% using 33.77 to 45.37 thousand tokens. **Independent selection yields substantial gains, and iterative search achieves higher detection rates with greater generation expenditure.**

Among the iterative instantiations, proposal generation accounts for most output-token variation under their shared candidate budget, linking expenditure to the work required to produce eligible rephrasings. Historical examples and reflective feedback also contribute to input-token consumption. Appendix A.4 reports reconstructed input tokens (Table 9), eligibility checks, and detector workloads (Table 10).

5 Analysis

We analyze the completed runs to assess semantic fidelity and examine how detection gains are distributed across sources.

5.1 Semantic Fidelity and Textual Properties

Across Gemma and Qwen discussions and Gemma research reviews, all four methods have high mean similarity to the original sources. Mean BERTScore F1 ranges from 0.9610 to 0.9888 and mean MPNet similarity from 0.8849 to 0.9835. On greedy Gemma discussions, the additional MiniLM means range from 0.8807 to 0.8981. MPNet and MiniLM provide independent diagnostics alongside the BERTScore eligibility check. Table 26 reports means, lower-tail similarities, and length ratios; Table 27 shows that formatting normalization has little effect on aggregate similarities.

A stratified manual audit judged 44 of 48 discussion pairs and all 16 research pairs to preserve the original meaning. Within these meaning-preserving subsets, TPR at the fixed Binoculars threshold increases by 15.91 percentage points for discussions and 33.33 points for research reviews relative to the matched originals. Detection gains therefore also occur in rephrasings judged to retain the author's substantive content. Appendix E.2 describes the audit. Figure 2 provides a recorded source pair with AI-reply excerpts and detector scores.

To assess the role of length, we retain discussion pairs whose rephrased source word count and mean final-reply word count are each within 5% of their matched Original values. This restriction uses length alone. Under greedy Gemma generation, Hill-climb retains a TPR gain of 31.62 percentage points across 39 sources, and GEPA retains a gain of 20.18 points across 38 sources. Both 95% source-bootstrap intervals lie above zero. Substantial detection gains can therefore occur with nearly unchanged source and reply lengths. Table 28 reports confidence intervals and all conditions at 5% and 10% tolerances; Appendix E.3 also checks the effect of excluding unchanged sources.

5.2 Coverage and Fixed-Threshold Detection

We focus on the Gemma experiments, which cover both ChangeMyView discussions and AI-Researcher reviews. To determine whether the aggregate gains are concentrated in a small number of sources, we examine source-level changes in detection counts and reply-score distributions.

For Hill-climb under greedy generation, the number of detected replies increases for 234 of 347 discussion posts and decreases for 23, producing 498 additional detections overall. The aggregate gain therefore reflects improved detection across a majority of evaluated posts. The median reply score decreases from 0.8750 to 0.8212 on discussions and from 0.8756 to 0.8473 on research reviews. In both tasks, the median moves from above to below the fixed Binoculars threshold of approximately 0.8536. These distributional shifts accompany the large fixed-threshold TPR gains.

Under sampled generation, the gains are less broadly distributed across sources. For Hill-climb on Gemma discussions, detection counts increase for 154 of 347 posts and decrease for 70, compared with 234 increases and 23 decreases under greedy generation. Net additional detections fall from 498 under greedy decoding to 143 under sampling. The smaller sampled TPR gain accompanies improvements on fewer posts and reduced detection on more posts relative to Original. Table 25 provides the complete count decomposition across methods, models, and tasks.

6 Conclusion

We introduced REAIR, a framework for improving the detection of subsequent AI replies through meaning-preserving rephrasing of an author's source text. REAIR formulates source preparation as constrained optimization using feedback from an existing detector, with four search instantiations demonstrating substantial gains across discussion and research-review tasks. Automatic similarity measures and human assessment support the semantic fidelity of selected rephrasings, and gains persist when source and reply lengths change little. Cross-detector results further show that the effect can extend beyond the detector used for selection. REAIR therefore gives existing detectors an active role before reply generation, complementing post hoc detection without modifying the detector or requiring the reply model to produce a watermark. Limited transfer across reply models and prompts motivates source-selection methods that remain effective across broader reply environments.

AI USE STATEMENT

Generative AI models produced the source rephrasings and replies evaluated in this study, as described in Sections 3 and 4. Generative AI tools also assisted with code implementation, manuscript drafting, and language editing. The authors reviewed the AI-assisted work and take responsibility for the final manuscript, code, claims, and reported results.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the search procedures, candidate budgets, generation settings, eligibility rules, detector threshold, and final evaluation protocol. Appendix A provides data and model configurations, complete prompts, update rules, query and token accounting, and statistical procedures. Appendices B through G provide the complete results and additional evaluation protocols. Final evaluation replies are generated after source selection and excluded from search decisions.

ETHICS STATEMENT

This study concerns authors rephrasing their own text to assess subsequent AI-generated replies. Such rephrasing should preserve the author's substantive meaning. Detector outputs are imperfect evidence: the observed human-reference false-positive rates are reported in Table 12, and a detection should not be used as the sole basis for accusations or punitive decisions. Reuse of the source datasets should respect their terms and contributors' privacy. The manual semantic audit described in Appendix E.2 assesses preservation of source meaning. Our detection evaluation uses human responses to original sources and does not establish how rephrasing affects new human responses.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6b8c6f846c3575e1d1ad496abea28826-Abstract-Conference.html.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634: 818–823, 2024. doi: 10.1038/s41586-024-08025-4.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.674. URL https://aclanthology.org/2024.acl-long.674/.

Google DeepMind. Gemma 4. Model release and documentation, 2026. URL https://deepmind.google/models/gemma/gemma-4/.

Google Workspace. Gemini in Gmail Will Now Provide Contextual Smart Replies. Google Workspace Updates, September 2024. URL https://workspaceupdates.googleblog.com/2024/09/contextual-smart-replies.html.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537, 2024. URL https://proceedings.mlr.press/v235/hans24a.html.

ICML. ICML 2026 Policy for LLM Use in Reviewing, 2026. URL https://icml.cc/Conferences/2026/LLM-Policy.

Rajesh Jayaram, Vincent Cohen-Addad, Alekh Agarwal, Miroslav Dudík, Sharon Li, and Martin Jaggi. ICML Experimental Program Using Google's Paper Assistant Tool (PAT). ICML Blog, January 2026. URL https://blog.icml.cc/2026/01/14/icml-experimental-program-using-googles-paper-assistant-tool-pat/.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, volume 202, pp. 17061–17084, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4-Abstract-Conference.html.

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Waterfall: Scalable framework for robust text watermarking and provenance for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20432–20466. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1138. URL https://aclanthology.org/2024.emnlp-main.1138/.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *International Conference on Machine Learning*, volume 235, pp. 29575–29620, 2024. URL https://proceedings.mlr.press/v235/liang24b.html.

Yepeng Liu, Tianyi Chen, Xuandong Zhao, Dawn Song, and Yuheng Bu. Learning to follow in-context watermark instructions via self-distillation, 2026a. URL https://arxiv.org/abs/2608.29030.

Yepeng Liu, Xuandong Zhao, Christopher Kruegel, Dawn Song, and Yuheng Bu. In-context watermarks for large language models. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/571082ea18d30060177dfcaf662ff0e5-Abstract-Conference.html.

Ning Lu, Shengcai Liu, Rui He, Yew-Soon Ong, Qi Wang, and Ke Tang. Large language models can be guided to evade AI-generated text detection. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=lLE0mWzUrr.

Haitong Luo, Weiyao Zhang, Suhang Wang, Wenji Zou, Chungang Lin, Xuying Meng, and Yujun Zhang. SpecDetect: Simple, fast, and training-free detection of LLM-generated text via spectral analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32356–32364, 2026. doi: 10.1609/aaai.v40i38.40510.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, volume 202, pp. 24950–24962, 2023. URL https://proceedings.mlr.press/v202/mitchell23a.html.

Choonghyun Park, Hyuhng Joon Kim, Junyeob Kim, Youna Kim, Taeuk Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-goo Lee, and Kang Min Yoo. Investigating the influence of prompt-specific shortcuts in AI generated text detection. *arXiv preprint arXiv:2406.16275*, 2024. URL https://arxiv.org/abs/2406.16275.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.18290.

Vishisht Srihari Rao, Aounon Kumar, Himabindu Lakkaraju, and Nihar B. Shah. Detecting LLM-generated peer reviews. *PLOS ONE*, 20(9):e0331871, 2025. doi: 10.1371/journal.pone.0331871.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0664. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2567c95fd41459a98a73ba893775d22a-Abstract-Conference.html.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 613–624, 2016. doi: 10.1145/2872427.2883081.

A. M. Turing. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, October 1950. doi: 10.1093/mind/LIX.236.433. URL https://academic.oup.com/mind/article/LIX/236/433/986238.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free LLM-generated text detection by mining token probability sequences. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06072.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan

Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Xiaowei Zhu, Yubing Ren, Fang Fang, Qingfeng Tan, Shi Wang, and Yanan Cao. DNA-DetectLLM: Unveiling AI-generated text via a DNA-inspired mutation-repair paradigm. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/0287c393907259ae6a269fca5e3509cd-Abstract-Conference.html.

A Detailed Experimental Protocol

This appendix specifies the procedures used for the results in the main text. Appendices B–D provide the complete detection and transfer results; Appendix E supports the source-level and semantic analyses. Appendices F and G specify the watermark comparison and preference-training experiments.

A.1 Data and Model Configurations

Data and task instructions. ChangeMyView contains authors' viewpoints and invites replies that challenge or refine their reasoning (Tan et al., 2016). Each of the 347 English posts has one human reference reply. The 49 AI-Researcher proposals describe natural language processing research problems, methods, and experimental plans; their presentation was standardized with an LLM in the original study (Si et al., 2025). Each proposal has one to three human reviews, giving 118 references. The reply model assesses scientific merit and feasibility through scores and written justifications. In both tasks, the original or rephrased source is supplied directly with the task instruction. The tasks are evaluated separately using their existing human responses as detection references.

Generation controls. Source rephrasing uses temperature 0.7. Reply generation uses greedy decoding at temperature 0 or stochastic decoding at 0.7. Methods share the reply instruction and decoding settings within each model pair, task, and reply temperature. Rephrasing prompts specify each method's proposal strategy and feedback context (Appendix A.10).

Table 6: Datasets and model combinations represented in the evaluation.

Data	Rephrasing model	Reply model	Task	Sources	Human responses
ChangeMyView	Gemma 4 E4B IT	Gemma 4 31B IT	discussion	347	347
ChangeMyView	Qwen3-8B	Qwen3-32B	discussion	347	347
ChangeMyView	DeepSeek V4 Flash	DeepSeek V4 Flash	discussion	347	347
Research proposals	Gemma 4 E4B IT	Gemma 4 31B IT	review	49	118

IT denotes instruction-tuned models.

Table 7: Generation and search settings for Gemma discussions.

Setting	Rephrasing generation	Reply generation
Temperature	0.7	0 or 0.7
Nucleus probability	0.8	0.8 when sampled
Top-k sampling	20	20 when sampled
Repetition penalty	1.05	1.00
Maximum new tokens	3,072	512
Candidates per round	3	
Final replies per source		3
Explicit reasoning trace	disabled	disabled

Qwen research-review screening. The preliminary screening evaluated Qwen3-32B on the 49 original proposals using the research-review instruction. Three reviews per proposal give 147 generated reviews at each temperature. Binoculars uses the 1,024-token observation window and the published threshold specified for this task, with the same 118 human references. Table 8 reports full detection at both temperatures, motivating the focus on Gemma for research-review optimization.

Table 8: Preliminary detection of Qwen3-32B reviews of original research proposals. TPR uses the fixed published Binoculars threshold.

Temperature	Detected reviews	TPR %	AUROC
0	147 / 147	100.00	0.9908
0.7	147 / 147	100.00	0.9808

A.2 Whole-Document BERTScore Implementation

The gate is applied after each rephrasing is proposed and before any selection replies are generated. It always compares the candidate with the fixed original source, including when the proposal input is the current best rephrasing. Candidates must satisfy both the BERTScore F1 threshold of 0.95 and the word-count ratio range of 0.50 to 1.40. These criteria define the automatic eligibility gate. Human assessment is reported in Appendix E.2.

The semantic gate tokenizes the complete source and rephrasing with RoBERTa-large and divides each token sequence into deterministic, non-overlapping windows of 510 content tokens. Each window is embedded separately with its own boundary tokens. The content-token embeddings are then concatenated, and ordinary BERTScore greedy cosine matching is performed over all source and rephrasing tokens. Precision averages each rephrasing token's best match over the full source; recall averages each source token's best match over the full rephrasing; their harmonic mean is the reported document F1. Matching may therefore cross window boundaries after embedding, and no source or rephrasing token is truncated.

Meaning preservation and factual anchors. The similarity threshold restricts candidates to reformulations closely aligned with the source while allowing changes in wording, syntax, and organization. The word-count bounds limit compression or expansion that could change the amount of information conveyed. Differences in numerical values, dates, URLs, and citation identifiers are recorded for manual inspection and do not independently determine eligibility. For research proposals, experiment-defining materials and values are protected and preserved verbatim. The research rephrasing prompt in Appendix A.10 specifies how these spans are marked and restored.

A.3 Search and Evaluation Budgets

BoN evaluates eight candidates with three replies each, for 24 selection replies per source. Each iterative method runs ten rounds with three candidate slots and three initial replies per eligible candidate, allocating 30 candidate slots and up to 90 initial selection replies. A challenger receives three additional replies only when its initial mean is strictly lower than the stored six-reply mean of the

current best source. The Original source receives six stored search replies at initialization. If all 30 candidate slots produce eligible rephrasings, an iterative search therefore uses between 96 and 126 reply generations for selection: six at initialization, 90 for initial candidate evaluation, and zero to 30 for challenger confirmation. The final three evaluation replies are separate from every search reply. Rephrasing proposal and GEPA reflection calls are counted separately from reply generation.

The three iterative methods share the candidate budget and final evaluation count. Their difference lies in proposal generation. Hill-climb conditions on the current best text and its scores. Hill-climb-ICL rephrases the current best source at every round and conditions on good and bad trajectory examples from that source, with at most half of the example budget assigned to each group. The fixed original source remains the semantic-preservation reference. GEPA reflects on a concrete source rephrasing, generated replies, detector scores, and eligibility diagnostics.

In the Gemma and Qwen discussion experiments, each requested rephrasing allows at most 100 generation attempts and each requested reply allows at most six. The research experiments allow nine attempts per rephrasing, consisting of one initial proposal and up to eight attempts with the retry instruction in Appendix A.10, and six attempts per reply. All attempts use the same eligibility criteria. An exhausted iterative proposal slot is recorded as ineligible and receives no selection replies. Such a slot uses part of the fixed candidate budget. A candidate requires three valid scored selection replies to enter the comparison, and a challenger requires three additional valid scored replies for acceptance. An incomplete comparison retains the incumbent. BoN requires its specified number of eligible candidates. Completed final evaluations require all three valid scored replies for every source. Unresolved required outputs are retained as incomplete until recovery under the same protocol.

Independent final evaluation. After selection, each chosen source is fixed and used to generate three new evaluation replies. Their random seeds do not overlap with those used during search, and the replies are excluded from all search decisions. Original receives the same three-reply evaluation under the same reply model, instruction, and temperature. New generation calls under greedy decoding can reproduce the same response.

A.4 Generation and Scoring Workload

Empirical generation cost is measured on the completed Gemma runs for the 347 discussion sources at each reply temperature. For each method, we sum the recorded number of generated tokens over all proposal and reply attempts in the search. This includes eligibility rejections, replacement proposals, and reply-generation retries. The proposal column contains rephrasing-model outputs, including GEPA's reflective proposal outputs. The search-reply column contains initialization replies, initial candidate replies, and additional challenger replies. The final-evaluation column contains the three separately generated replies per selected source and any recorded retries needed to obtain them. Each condition has 1,041 retained final replies.

Counts use the generating model's tokenizer and include a terminal end-of-sequence token when one is generated. Padding is excluded. The Gemma discussion totals are divided by 347 and expressed in thousands of tokens. Development searches, smoke tests, Original control evaluation, and intermediate evaluation at rounds three and five are outside this accounting. Table 5 reports output tokens. Table 9 adds input tokens reconstructed from the saved requests, including the complete source, historical examples, and reflective feedback. Each prompt is checked against its recorded hash and rendered with the corresponding model's chat template and tokenizer. Each recorded attempt contributes its input tokens, including repeated prompts on retries. The reconstruction covers 417,865 attempts across the eight conditions. Input counts exclude padding. Proposal prompts account for 113.12 to 128.28 thousand input tokens per source for Hill-climb, 131.01 to 153.80 thousand for Hill-climb-ICL, and 518.81 to 614.44 thousand for GEPA. The history and feedback used to construct proposals therefore contribute substantially to input processing.

Table 10 reports the recorded checking workload. Every proposal attempt receives the original-referenced eligibility assessment, including attempts that are rejected. RoBERTa windows count the source and candidate windows processed by these assessments. Detector counts refer to completed text scores during search and final evaluation. Binoculars uses both its observer and performer for

Table 9: Reconstructed input-token consumption for Gemma on ChangeMyView, in thousands of tokens per source. Search total sums proposal and search-reply inputs. Final evaluation is accounted for separately. Counts include all recorded attempts and exclude padding.

Method	Temp.	Proposal	Search reply	Search total	Final eval.
BoN	0	3.90	9.38	13.28	1.18
Hill-climb	0	113.12	38.96	152.07	1.14
Hill-climb-ICL	0	153.80	39.70	193.50	1.14
GEPA	0	614.44	38.60	653.04	1.11
BoN	0.7	3.78	9.41	13.19	1.17
Hill-climb	0.7	128.28	39.51	167.79	1.15
Hill-climb-ICL	0.7	131.01	40.91	171.93	1.15
GEPA	0.7	518.81	39.93	558.74	1.14

Table 10: Recorded eligibility and detection work per source for Gemma on ChangeMyView. Eligibility checks include rejected proposals and retries. RoBERTa windows count encoder inputs. Detector columns count completed text scores. Each condition contains 347 sources.

		Eligibility	RoBERTa	Detector scores	
Method	Temp.	checks	windows	Search	Final eval.
BoN	0	8.98	21.00	24.00	3.00
Hill-climb	0	104.63	226.48	102.47	3.00
Hill-climb-ICL	0	55.43	123.05	104.59	3.00
GEPA	0	88.86	196.92	102.81	3.00
BoN	0.7	8.61	20.21	24.00	3.00
Hill-climb	0.7	122.43	261.56	104.09	3.00
Hill-climb-ICL	0.7	43.66	98.15	107.04	3.00
GEPA	0.7	71.55	160.55	105.71	3.00

each score. Initialization is included once within each independent search, and final evaluation is kept separate. These counts measure the recorded workload separately from generated-token counts.

A.5 Independent Candidate Selection

BoN applies the base rephrasing prompt to the original source for each proposal. Candidates pass the same original-referenced eligibility checks used by the iterative instantiations. Each of the N eligible candidates receives three selection replies, and their mean Binoculars scores determine the selected candidate. The lowest mean is selected, with output order resolving a tie. The selected source receives three separate final replies. The main setting uses $N = 8$; the additional greedy Qwen comparison uses $N = 30$.

A.6 Exact Challenger and Acceptance Rules

Let x denote the original source and b_r the current best source before round r. Initialization sets $b_1 = x$ and generates six replies to x. Their mean Binoculars score is stored as $\bar{s}_6(b_1)$. The original source and its six scores remain available throughout the search.

Before a candidate can become eligible, it must pass the original-referenced semantic and length checks described above. Each eligible candidate c receives three initial selection replies, with mean $\bar{s}_3(c)$. A candidate qualifies to challenge the current best only when $\bar{s}_3(c) < \bar{s}_6(b_r)$. Among qualifying candidates, the candidate with the lowest initial mean becomes the challenger. Equal candidate means are resolved by candidate output order, selecting the earlier slot. If no candidate qualifies, the current best is retained and no additional confirmation replies are requested in that round.

The challenger receives three additional replies. Its comparison score is the equally weighted mean of all six replies: the three initial selection replies and the three additional confirmation replies.

Writing their respective means as $\bar{s}_{\text{selection}}(c)$ and $\bar{s}_{\text{confirmation}}(c)$ gives

$$\bar{s}_6(c) = \frac{\bar{s}_{\text{selection}}(c) + \bar{s}_{\text{confirmation}}(c)}{2}. \tag{2}$$

The challenger replaces the current best exactly when $\bar{s}_6(c) < \bar{s}_6(b_r)$. Equality retains the current best. An accepted candidate carries its six replies and their mean into subsequent rounds. Comparisons reuse these stored observations; they do not generate another set of replies for the incumbent. If no candidate is accepted during the search, the original source remains the final source.

After ten rounds, the selected source is fixed and receives three separately generated final replies. Initialization, initial selection, additional confirmation, and final evaluation use separate generation records and seed namespaces. Final evaluation replies never determine acceptance or historical-example selection. At temperature zero, repeated calls may return identical text; the protocol separates search from evaluation without assuming three distinct sampled responses. All statistical comparisons retain the source as the unit of inference.

A.7 Construction of Historical Examples

Hill-climb-ICL builds its history separately for each source and each run. Every eligible candidate with three scored selection replies enters that source's history, including candidates that were not accepted. The first round has no historical examples. At evaluation time, examples come from the current evaluation source's own earlier candidates. No example is borrowed from another source, another method, or a separately conducted run.

Good examples first include candidates that previously became the current best, ordered from the most recently accepted candidate backward. Remaining good-example positions are filled by candidates with the lowest initial three-reply selection means. Bad examples are candidates with the highest initial selection means after excluding every historical current-best candidate and all selected good examples. Bad examples receive no additional evaluation for inclusion in the prompt. The displayed score of every example is its initial three-reply selection mean, including examples that later received confirmation replies.

Let N be the smaller of K and the number of eligible historical candidates. The requested group sizes are $\min(K/2, \lceil N/2 \rceil)$ good examples and $\min(K/2, \lfloor N/2 \rfloor)$ bad examples. Any unfilled positions remain empty when the exclusion rules leave too few candidates. Equal scores in the good-example ranking are resolved by earlier round and then a stable candidate identifier. Equal scores in the bad-example ranking are resolved by later round and then the identifier. The same candidate record cannot appear in both groups.

A.8 Number of Historical Examples

The development comparison evaluates $K = 4$ and $K = 8$ on a fixed set of 50 discussion posts drawn from the 2,250-post training collection. Posts are ordered by the SHA-256 digest of their source identifiers, and the first 50 are selected. This development set is separate from the 347-post evaluation collection. Each setting runs all ten search rounds and receives three final replies per development source, yielding 150 final replies per setting.

Selection uses the mean Binoculars score of these final replies. If the absolute difference between the two means is below 0.001, the rule selects $K = 4$. Otherwise, it selects the setting with the lower mean. Table 11 gives all six development comparisons. None falls within the 0.001 tolerance. The selected value is fixed before the corresponding 347-post evaluation. The research-review experiments inherit the Gemma discussion setting, $K = 8$ at both reply temperatures. The 49 research proposals are not used to select K.

Table 11: Historical-example selection on 50 development posts. Values are final three-reply mean Binoculars scores averaged across sources. Lower values are preferred.

Model pair	Reply temperature	$K = 4$	$K = 8$	Selected K
Gemma	0	0.822752	0.820582	8
Gemma	0.7	0.861116	0.858916	8
Qwen	0	0.788363	0.780337	8
Qwen	0.7	0.838360	0.836215	8
DeepSeek	0	0.921456	0.931834	4
DeepSeek	0.7	0.946376	0.951031	4

For DeepSeek at temperature zero, the development searches retain 13 and 30 generation records that exhausted their retry limits for $K = 4$ and $K = 8$, respectively. Each setting has 150 valid final evaluation replies.

A.9 Source-Specific Use of GEPA

GEPA optimizes one concrete source text at a time. The optimized object is the current rephrasing of that source. The reflection instruction remains fixed throughout the experiment. We use the GEPA library's stateless reflective-mutation proposer within a source-specific controller with ten rounds and three candidate slots per round. The controller implements the challenger and acceptance rules in Appendix A.6.

At initialization, the proposal model receives the original source, its six generated replies and detector scores, and the corresponding semantic and length diagnostics. In each later round, it receives the current best rephrasing and feedback records from the preceding round. These records include the fixed original source, proposed rephrasings, available reply texts and scores, initial score means, semantic-preservation and length diagnostics, and acceptance outcomes. For a challenger, the feedback also includes its six-reply comparison score. An ineligible candidate contributes its rejection diagnostics without downstream reply scores.

Three proposals are requested from the same current rephrasing and feedback context in each round. Each proposal is checked against the fixed original source. Eligible proposals receive three selection replies, and the qualifying challenger receives three additional replies. The controller accepts a proposal only when its pooled six-reply mean improves on the incumbent's stored mean. The accepted source becomes the starting text for the next round. Reflection and proposal generation therefore revise the source itself; no reusable instruction is learned or transferred between sources.

A.10 Complete Rephrasing and Response Prompts

The following transcripts preserve the prompt wording. Display line breaks are adjusted for page width. Braced placeholders are replaced with the indicated source text, candidate text, or observed score. Prompt delimiters and output-field names are reproduced literally.

Discussion rephrasing prompt. BoN inserts the original source into this template. Hill-climb and Hill-climb-ICL both insert the current best rephrasing; the original source remains the fixed semantic-preservation reference. Hill-climb and Hill-climb-ICL each append their context block after the base prompt, separated by a line containing three hyphens and a blank line on each side.

```
Rewrite the text below in a natural and coherent way while preserving
its original meaning.

The rewrite should express the same ideas, viewpoint, and reasoning as
the original without introducing any substantive change. You may freely
vary the wording, sentence structure, and organization.

Return only the rewritten text. Do not include analysis, planning, JSON,
labels, markdown, or comments about the task.
```

```
Original text:
{POST_TEXT}

Rewrite:
```

Hill-climb. The context block below accompanies the base rephrasing prompt. The original source remains the fixed reference for meaning preservation. Both displayed means use the stored six-reply observations.

```
Optimization context from the previous round:

Original text:
<<<
{ORIGINAL_TEXT}
>>>

Current best rewrite:
<<<
{CURRENT_BEST_REWRITE}
>>>

Mean Binoculars score of replies to the original text:
{ORIGINAL_REPLY_MEAN_SCORE}
Mean Binoculars score of replies to the current best rewrite:
{CURRENT_BEST_REPLY_MEAN_SCORE}

Lower Binoculars score means the downstream replies are more AI-like.
Produce one new rewrite of the current best rewrite that may improve it
while preserving the meaning of the original text. The current best
rewrite is the text to rewrite; the original text is only the fixed
meaning-preservation reference. Do not merely copy the current best
rewrite. Follow the output rules in the base rewrite prompt.
```

Hill-climb-ICL. The discussion experiments use the same rephrasing template with the current best rephrasing as its input and the following context block. The original source is used only as the fixed semantic-preservation reference. The original and current-best means again use six replies. Historical-example scores use the initial three replies, as specified in Appendix A.7.

```
Optimization context from the previous rounds:

Original text:
<<<
{ORIGINAL_TEXT}
>>>

Current best rewrite:
<<<
{CURRENT_BEST_REWRITE}
>>>

Mean Binoculars score of replies to the original text:
{ORIGINAL_REPLY_MEAN_SCORE}
Mean Binoculars score of replies to the current best rewrite:
{CURRENT_BEST_REPLY_MEAN_SCORE}

Lower Binoculars score means the downstream replies are more AI-like.
The examples below are valid rewrites generated earlier for this same
original text. Good examples had lower selection means; bad examples had
higher selection means. Use them only as evidence about effective and
ineffective rewriting choices.

Historical examples:
{EXAMPLES}
```

```
Produce one new rewrite of the current best rewrite that may improve
on it while preserving the original meaning. Use the current best
rewrite as the starting text for this round. Do not copy an example.
Follow the output rules in the base rewrite prompt and output only the
rewrite text.
```

Good examples are listed before bad examples. Each example uses the following format, with its selection mean displayed to ten decimal places. Blocks are separated by a blank line. With no eligible history, the example block contains exactly the sentence shown beneath the example formats.

```
Label: good
Round: {ROUND}
Selection mean: {SELECTION_MEAN}
Rewrite:
<<<
{CANDIDATE_TEXT}
>>>

Label: bad
Round: {ROUND}
Selection mean: {SELECTION_MEAN}
Rewrite:
<<<
{CANDIDATE_TEXT}
>>>

No valid historical examples are available yet.
```

GEPA reflection prompt. The current-parameter placeholder contains the current concrete rephrasing. The feedback placeholder contains the evaluation records described in Appendix A.9. The following instruction directly requests the next source rephrasing.

```
You are the GEPA reflection and proposal model for one source-specific
direct-rewrite search.

The artifact being optimized is one concrete rewrite, not a reusable
prompt or instruction.

Current candidate artifact:
<<<
<curr_param>
>>>

The evaluator feedback below contains the fixed original source, actual
downstream replies, raw Binoculars scores, semantic-preservation
diagnostics, length diagnostics, and outcomes from the preceding search
round:

<side_info>

Propose one new version of the concrete rewrite. Preserve the original
source's meaning, viewpoint, and reasoning without introducing any
substantive change. Express the same ideas naturally and coherently
while varying wording, sentence structure, or organization where useful.
Use the evaluator feedback to improve the rewrite, but do not copy reply
text and do not produce a general strategy or prompt.

Return only the rewritten text. Do not include analysis, planning, JSON,
labels, markdown fences, comments, or any description of the task.
```

Research-proposal rephrasing prompt. Research proposals use the following task-specific base prompt. Protected spans identify experiment-defining materials and values. The proposal is pre-

sented with these spans marked, and their original contents are restored before response generation. Hill-climb and Hill-climb-ICL both rephrase the marked current best proposal, while semantic checks remain anchored to the marked original proposal. The in-context blocks above remain unchanged and precede the research rephrasing prompt, separated by a line containing three hyphens and a blank line on each side. Research GEPA uses the reflection prompt in the same preceding position and supplies the marked current proposal to the research rephrasing template.

```
Rewrite the academic research proposal below in a natural and coherent
academic style while preserving its meaning. Keep the same research
question, motivation, method, experimental design, examples, limitations,
qualifications, and fallback strategy. Do not improve, correct, extend,
or otherwise change the research idea.

The proposal contains <LOCKED id="L0001">...</LOCKED> blocks. Their
contents are experiment-defining materials or values. Copy every LOCKED
block exactly, including its opening tag, contents, and closing tag,
exactly once and in the same order. The locked text is visible context
for rewriting the surrounding text. Rewrite the text outside the LOCKED
blocks naturally while preserving its meaning.

Rewrite the unlocked prose throughout the proposal; do not return it
unchanged.

Return one complete proposal using exactly this plain-text structure:

Title: <title>

1. Problem Statement:
<problem statement>

2. Motivation:
<motivation>

3. Proposed Method:
<proposed method>

4. Step-by-Step Experiment Plan:
<experiment plan>

5. Test Case Examples:
<test case examples>

6. Fallback Plan:
<fallback plan>

Keep all information from the original proposal. Preserve the ordering
and dependencies of the method and experiments. Do not add evidence,
results, citations, datasets, methods, examples, or implementation
details. Do not remove qualifications, uncertainties, limitations, or
fallback options. Do not strengthen or weaken the claims. If information
is missing, do not supply it.

Keep meaningful subheadings, bullets, numbered steps, and multi-stage
examples when they exist. The title may be rephrased only if its full
meaning is preserved. Do not add, remove, merge, rename, or reorder the
six main sections.

Return only the complete rewritten proposal with the unchanged LOCKED
blocks. Do not include analysis, planning, JSON, Markdown fences, a
preface, or comments about the rewriting process.

Original research proposal:
<<<
{POST_TEXT}
```

```
>>>

Rewritten research proposal:
```

Subsequent generation attempts use the same template with the instruction governing edits to unlocked prose replaced by the following paragraph. The rest of the prompt remains unchanged.

```
Your previous output was identical to the input. Rewrite the editable
prose in different wording while preserving the same meaning. Keep every
LOCKED block exactly unchanged.
```

A.10.1 Discussion Reply Prompts

The discussion experiments use the two reply prompts below. Direct Rebuttal is used during source selection and matched evaluation. Prompt transfer retains the selected sources and changes the reply prompt. Model transfer retains Direct Rebuttal and changes the reply model. Within each evaluation condition, Original and all evaluated rephrasings receive the same reply prompt. The placeholder {POST_TEXT} contains the original or selected source.

Reply prompt	Experimental role
Direct Rebuttal	Source selection, matched evaluation, and reply model transfer.
Concession and Rebuttal	Reply prompt transfer with Gemma 4 31B and Qwen3-32B.

Direct Rebuttal

```
The post presents someone's opinion and reasoning.

Write a short, thoughtful reply that directly responds to the author's
view. Your goal is to respectfully challenge or refine their argument in
a way that could plausibly change their mind. Keep the reply natural,
clear, and concise.

Post:
<<<
{POST_TEXT}
>>>

Reply:
```

Concession and Rebuttal

```
Help me write a response to the r/ChangeMyView post below. Start by
acknowledging the strongest or most reasonable part of the author's
position. Then explain why the overall conclusion still does not
necessarily follow and present a more convincing alternative view.
Keep the reply concise and conversational. Use no more than 120 words.

Post:

---
{POST_TEXT}

Reply:
```

A.10.2 Research Review Prompt

The research-feedback experiment uses the complete review instruction below for Original and rephrased proposals.

```
You are an expert reviewer in Natural Language Processing, Machine
```

```
Learning, and Large Language Models. Review the research proposal below.
Write a thoughtful, direct, natural, clear, and concise expert review
that evaluates the proposal itself.

Evaluate the following dimensions using the AI-Researcher study criteria:

Novelty Score (1-10): Judge whether the idea is creative, meaningfully
different from existing work, and offers fresh insight. Treat work
publicly available before July 2024 as existing work. A score of 1 means
clearly duplicated by many existing ideas; 5 means somewhat different
but probably not enough for a new paper; 6 means reasonably novel with
notable differences; 8 means clearly novel; and 10 means very novel in
an interesting and clever way.

Feasibility Score (1-10): Judge whether a typical computer-science PhD
student could implement and evaluate the idea within one to two months,
assuming abundant OpenAI/Anthropic API access but limited GPU compute.
A score of 1 means impossible or fundamentally flawed; 3 means very
challenging or beyond normal academic resources; 5 means moderately
feasible with careful planning or proposal changes; 6 means feasible;
8 means highly feasible; and 10 means executable within a few days.

Expected Effectiveness Score (1-10): Judge how likely the method is to
work well and improve over the stated baselines. A score of 1 means
extremely unlikely due to major flaws; 3 means likely to work only in
special cases; 5 means improvements would probably be marginal or
inconsistent; 6 means moderate gains are plausible on some benchmarks;
8 means significant improvements are probable; and 10 means large
improvements across many benchmarks are highly likely.

Excitement Score (1-10): Judge the potential scientific impact if the
project is fully executed. A score of 1 means no identifiable or
interesting contribution; 3 means marginal and incremental; 5 means
interesting in parts but not sufficiently exciting; 6 means suitable
for a major conference but still limited or incremental; 8 means likely
to deepen understanding or make major progress; and 10 means
transformative and best-paper caliber.

Overall Score (1-10): Use the major-conference scale: 1 critically
flawed, trivial, or wrong; 2 strong reject; 3 clear reject; 4 reject but
potentially useful; 5 marginally below acceptance; 6 marginally above
acceptance; 7 accept; 8 clear accept/top 50%; 9 strong accept/top 15%;
10 seminal/top 5%.

For every score, provide a proposal-specific rationale of at least two
to three complete sentences. Explain concrete strengths and weaknesses
instead of merely restating the score. For a low novelty score, identify
similar lines of work when you can do so reliably. For a low feasibility
score, identify the difficult components and why they are difficult.
Keep the five rationales distinct rather than repeating the same
criticism.

Finish with Confidence (1-5): 1 is an educated guess; 2 means important
parts may have been misunderstood; 3 means fairly confident; 4 means
confident but not absolutely certain; and 5 means absolutely certain
and very familiar with the relevant literature.

Use exactly this output field order:

Novelty Score: <integer 1-10>
Novelty Rationale: <at least 2-3 sentences>

Feasibility Score: <integer 1-10>
Feasibility Rationale: <at least 2-3 sentences>
```

```
Expected Effectiveness Score: <integer 1-10>
Expected Effectiveness Rationale: <at least 2-3 sentences>

Excitement Score: <integer 1-10>
Excitement Rationale: <at least 2-3 sentences>

Overall Score: <integer 1-10>
Overall Rationale: <at least 2-3 sentences>

Confidence: <integer 1-5>

Do not output administrative, identity, or attestation fields: Name,
Institution, Affiliation, Email, Consent, Honor Code, Familiarity,
Reviewing Experience, Timestamp, or Time Spent. Do not output personal
information. Return only the review fields above, without analysis,
planning, JSON, markdown fences, a preface, or a closing comment.

Research proposal:
<<<
{POST_TEXT}
>>>

Review:
```

A.11 Statistical Implementation

All paired comparisons use the original source as the unit of inference. For mean-score comparisons, each source contributes the mean of its three final reply scores. For TPR comparisons, each source contributes the fraction of its three final replies classified as AI-generated. Binoculars decisions use the fixed published threshold $\tau = 0.8536432310785527$. The threshold is retained throughout resampling.

Detector configurations. Binoculars follows its official implementation with Falcon-7B as the observer and Falcon-7B-Instruct as the performer (Hans et al., 2024). Lower scores indicate stronger evidence of AI generation. DNA-DetectLLM, LastDE Base, and SpecDetect Base use their respective published configurations (Zhu et al., 2025; Xu et al., 2025; Luo et al., 2026).

Metric-specific tests. The Gemma discussion, Qwen discussion, and research-review TPR comparisons use one-sided paired t-tests on source-level detection fractions. The alternative hypothesis is that the method increases detection relative to Original. The research-review value $p = 0.01187$ tests this TPR hypothesis for GEPA under sampled generation. Each instruction-transfer comparison uses Original generated with the same alternative instruction and temperature.

The Gemma and Qwen discussion cross-detector tables use one-sided paired t-tests on source-level mean scores. DNA-DetectLLM improvements correspond to lower scores; LastDE Base and SpecDetect Base improvements correspond to higher scores. Detector-specific selection comparisons use the same paired score test, with Binoculars-selected sources as the reference where indicated. The research-review cross-detector table uses one-sided paired sign-flip tests on source-level score improvements with 20,000 randomizations. Each randomization independently reverses the sign of each source's paired difference. Its Monte Carlo p-value is $(e + 1)/20{,}001$, where e is the number of randomized statistics at least as favorable as the observed mean improvement. Values near 5.00×10^{-5} therefore reflect this test's resolution. A score-test p-value refers to mean scores, independently of the AUROC shown in the same row. All reported p-values are unadjusted.

Confidence intervals and human references. Unless otherwise specified, 95% confidence intervals use 2,000 paired source-cluster bootstrap replicates. Detector score correlation and length analyses use 10,000 replicates. Resampling preserves the pairing between methods within each original source. AUROC is computed from individual generated and human response scores, oriented so that a higher ranking indicates stronger AI evidence. For discussion AUROC resampling, each sampled source brings its associated human reply and all generated replies into the replicate. The human reference collection is shared across compared methods and is resampled with the source

clusters. For research-review AUROC resampling, each sampled proposal brings all of its associated human reviews and all three generated reviews per method. The full reference collection contains 118 human reviews linked to 49 proposals; the number of human reviews in a bootstrap replicate can vary with the sampled proposals. Paired AUROC differences use 2,000 such replicates. The research cross-detector analysis additionally evaluates AUROC differences with 5,000 source-level paired swaps of the Original and method reply groups, keeping the human reference scores fixed during each permutation. The displayed research cross-detector p-values are the score sign-flip tests described above.

Detector input lengths. The ChangeMyView human-reference scores use a 512-token Binoculars window; 53 of the 347 human replies were truncated for scoring. The research-review runs use a 1,024-token window, and none of their 118 human references was truncated. The published decision threshold is fixed for both collections. Table 12 reports their observed FPRs under these preprocessing settings. Source-rephrasing similarity is computed over the complete documents as specified above.

Table 12 reports the false-positive rate obtained when the published low-FPR threshold is applied without recalibration to the two evaluation reference collections. A human response is counted as a false positive when its Binoculars score is strictly below τ. The human references are responses to the original sources and are shared across compared methods. Human responses to rephrased sources were not collected.

Table 12: Observed human-reference false-positive rates at the fixed published Binoculars low-FPR threshold $\tau = 0.8536432310785527$. The nominal 0.01% rate describes the detector's original calibration, not these evaluation collections.

Evaluation collection	Human responses	False positives	Observed FPR
ChangeMyView	347	7	2.02%
AI-Researcher	118	9	7.63%

B Complete Main Results

These tables expand the discussion results in Section 4.1 and the research-review results in Table 2, retaining both reply temperatures and all four search instantiations. TPR uses the same fixed published Binoculars threshold as in the main text.

B.1 Gemma Discussion Results

Table 13: Gemma discussion results on 347 posts at the fixed published Binoculars threshold. Delta is the TPR change from matched Original in percentage points. TPR p is the unadjusted one-sided paired t-test of increased source-level detection fractions.

Temperature	Method	TPR %	Delta	TPR p
0	BoN	60.90	31.51	3.78×10^{-25}
	Hill-climb	77.23	47.84	1.37×10^{-55}
	Hill-climb-ICL, $K=8$	**78.39**	**48.99**	3.99×10^{-61}
	GEPA	77.81	48.41	2.31×10^{-56}
0.7	BoN	37.37	7.20	4.75×10^{-4}
	Hill-climb	**43.90**	**13.74**	7.72×10^{-11}
	Hill-climb-ICL, $K=8$	43.13	12.97	2.46×10^{-9}
	GEPA	41.79	11.62	9.58×10^{-8}

B.2 QWEN DISCUSSION RESULTS

Table 14: Qwen discussion evaluation on 347 posts. TPR uses the fixed published Binoculars threshold; lower mean scores indicate stronger AI evidence. TPR p is the unadjusted one-sided paired t-test of increased source-level detection fractions relative to matched Original.

Temperature	Method	TPR %	Mean score	TPR p	AUROC	Delta AUROC
0	Original	65.42	0.828945		0.952930	0.000000
	BoN, $N = 8$	88.18	0.790115	5.36×10^{-21}	0.986813	0.033883
	BoN, $N = 30$	91.64	**0.775376**	3.30×10^{-26}	**0.992376**	**0.039446**
	Hill-climb	**92.12**	0.776057	3.10×10^{-30}	0.991511	0.038581
	Hill-climb-ICL, $K = 8$	89.91	0.786279	4.49×10^{-24}	0.991458	0.038528
	GEPA	91.16	0.777247	2.93×10^{-27}	0.991707	0.038778
0.7	Original	57.25	0.843094		0.940351	0.000000
	BoN, $N = 8$	63.50	0.837044	8.05×10^{-4}	0.945437	0.005087
	Hill-climb	63.78	0.834961	3.87×10^{-4}	**0.955027**	**0.014676**
	Hill-climb-ICL, $K = 8$	**66.47**	**0.833372**	1.07×10^{-6}	0.951623	0.011273
	GEPA	64.17	0.836234	1.54×10^{-4}	0.942992	0.002641

For the additional greedy Qwen BoN ($N = 30$) condition, TPR is 91.64%, mean Binoculars score is 0.775376, and AUROC is 0.992376. Hill-climb, which has the same nominal 30-candidate budget, reaches 92.12% TPR. Its paired TPR advantage is 0.48 percentage points with a 95% source-cluster confidence interval of [−2.40, 3.27] percentage points.

B.3 RESEARCH-PROPOSAL REVIEW RESULTS

Table 15: Gemma research-review evaluation on 49 proposals, with 147 final reviews per method and temperature. Delta is the fixed-threshold TPR change from matched Original in percentage points. TPR p is the unadjusted one-sided paired t-test of increased source-level detection fractions.

Temperature	Method	TPR %	Delta	Mean score	AUROC	TPR p
0	Original	12.93	0.00	0.878324	0.882884	
	BoN	40.82	27.89	0.860459	0.912833	4.48×10^{-6}
	Hill-climb	61.22	48.30	0.848619	0.931512	4.25×10^{-11}
	Hill-climb-ICL, $K = 8$	56.46	43.54	0.846668	0.932261	2.14×10^{-9}
	GEPA	**63.27**	**50.34**	**0.843935**	**0.933846**	1.18×10^{-10}
0.7	Original	11.56	0.00	0.879128	0.876110	
	BoN	17.01	5.44	0.876343	0.878041	0.1103
	Hill-climb	14.29	2.72	0.874004	**0.895394**	0.2553
	Hill-climb-ICL, $K = 8$	18.37	6.80	0.877299	0.873429	0.04795
	GEPA	**21.77**	**10.20**	**0.872803**	0.893434	0.01187

B.4 DEEPSEEK DISCUSSION RESULTS

Hill-climb-ICL uses $K = 4$ and rephrases the current best source in each round, with eligibility checks anchored to the original source. Each temperature has 347 sources and 1,041 valid final evaluation replies generated separately from search. The searches retain 171 and 2 generation records that exhausted their retry limits at temperatures 0 and 0.7, respectively. These records occur during candidate proposal, selection, or confirmation; the reported metrics use the valid final evaluation replies.

Table 16: DeepSeek V4 Flash discussion results on 347 posts, with 1,041 final replies per method and temperature. TPR uses the fixed published Binoculars threshold; lower mean scores indicate stronger AI evidence.

	Temperature 0			Temperature 0.7		
Method	TPR %	Mean score	AUROC	TPR %	Mean score	AUROC
Original	0.96	0.943917	0.643157	1.15	0.952066	0.595952
BoN	2.21	0.942414	0.647845	1.15	0.952921	0.589888
Hill-climb	2.59	0.938989	0.667143	**1.73**	**0.951836**	**0.596549**
Hill-climb-ICL, $K = 4$	2.69	0.936750	0.683437	1.15	0.952409	0.594573
GEPA	**2.88**	**0.936180**	**0.686213**	1.25	0.954465	0.580438

C Cross-Detector Evaluation

This appendix supports Section 4.3: it separates rescoring of Binoculars-selected replies from changing the selection detector on a fixed candidate pool.

C.1 Alignment with the Selection Detector

We compare paired score changes for Hill-climb and GEPA on the same 347 discussion sources. Each source contributes the mean of its three final replies. Improvements are oriented toward stronger AI evidence before computing Spearman correlations with the Binoculars improvement. Table 17 reports each model, temperature, and method separately. DNA-DetectLLM has the highest correlation in every comparison. Paired source bootstraps with 10,000 resamples also yield positive 95% intervals for its correlation difference from each other detector in all eight settings.

The scoring functions provide a structural connection. Let L be the mean negative log probability of observed tokens under the performer, M the corresponding value for its most probable tokens, and H the observer-to-performer cross entropy. With common model probabilities and observation windows, the implemented scores satisfy

$$s_{\text{Binoculars}} = \frac{L}{H}, \qquad s_{\text{DNA}} = \frac{L+M}{2H} = \frac{1}{2} s_{\text{Binoculars}} + \frac{M}{2H}. \tag{3}$$

Both detectors use Falcon-7B as observer and Falcon-7B-Instruct as performer in our experiments. Their configured precision and token windows are retained during evaluation. LastDE Base and SpecDetect Base use GPT-J-6B and different functions of token probabilities. The shared likelihood term and normalization explain the structural alignment of the first two scores. Table 17 measures how closely their changes agree on the evaluated replies.

Table 17: Spearman correlation of source-level score improvements with Binoculars. Each row includes 347 paired sources.

Temp.	Method	DNA-DetectLLM	LastDE Base	SpecDetect Base
Gemma discussions				
0	Hill-climb	0.938	0.303	0.145
0	GEPA	0.921	0.378	0.172
0.7	Hill-climb	0.928	0.300	0.165
0.7	GEPA	0.921	0.370	0.230
Qwen discussions				
0	Hill-climb	0.934	0.269	0.415
0	GEPA	0.946	0.313	0.418
0.7	Hill-climb	0.936	0.291	0.359
0.7	GEPA	0.943	0.304	0.447

C.2 GEMMA AND QWEN DISCUSSION RESULTS

Tables 18 and 19 evaluate the final discussion replies with three additional detectors. Rephrasing and selection use Binoculars throughout. The selected sources and final reply texts remain fixed during rescoring. Each detector compares the same 1,041 generated replies per method with its scores for the 347 human replies. DNA-DetectLLM assigns lower scores to stronger AI evidence. LastDE Base and SpecDetect Base assign higher scores to stronger AI evidence. Score p tests a favorable mean-score change from Original using the paired source-level procedure in Appendix A.11.

Table 18: Cross-detector evaluation of Gemma discussion replies selected using Binoculars. Lower DNA-DetectLLM scores and higher LastDE/SpecDetect scores indicate stronger AI evidence. Score p tests a favorable source-level mean-score change from matched Original, not an AUROC change.

Detector	Temperature	Method	Mean score	AUROC	Score p
DNA-DetectLLM	0	Original	0.617430	0.917337	
		BoN	0.596889	0.966439	2.45×10^{-33}
		Hill-climb	0.586277	0.977895	3.09×10^{-59}
		Hill-climb-ICL, $K=8$	0.584838	0.980118	1.05×10^{-58}
		GEPA	0.586836	0.979523	8.01×10^{-63}
	0.7	Original	0.617656	0.920573	
		BoN	0.610261	0.939501	6.15×10^{-9}
		Hill-climb	0.606855	0.946028	4.88×10^{-17}
		Hill-climb-ICL, $K=8$	0.607359	0.943656	9.62×10^{-16}
		GEPA	0.607462	0.948058	1.45×10^{-17}
LastDE Base	0	Original	–39.505826	0.678709	
		BoN	–39.175732	0.689497	0.05084
		Hill-climb	–38.634632	0.720456	1.26×10^{-6}
		Hill-climb-ICL, $K=8$	–38.545972	0.725076	2.44×10^{-7}
		GEPA	–39.252820	0.688919	0.08657
	0.7	Original	–39.754397	0.661509	
		BoN	–40.003472	0.649425	0.9397
		Hill-climb	–40.057026	0.640827	0.9729
		Hill-climb-ICL, $K=8$	–39.871012	0.651493	0.7818
		GEPA	–40.361855	0.622816	0.9999
SpecDetect Base	0	Original	–0.017505	0.460110	
		BoN	–0.017129	0.471593	0.05901
		Hill-climb	–0.017339	0.467800	0.2469
		Hill-climb-ICL, $K=8$	–0.016848	0.479936	0.002852
		GEPA	–0.017198	0.470975	0.1209
	0.7	Original	–0.017369	0.462687	
		BoN	–0.018053	0.449507	0.9994
		Hill-climb	–0.018197	0.447954	0.9999
		Hill-climb-ICL, $K=8$	–0.018004	0.449681	0.9995
		GEPA	–0.017727	0.456505	0.9679

Table 19: Cross-detector evaluation of Qwen discussion replies selected using Binoculars. Lower DNA-DetectLLM scores and higher LastDE/SpecDetect scores indicate stronger AI evidence. Score p tests a favorable source-level mean-score change from matched Original, not an AUROC change.

Detector	Temperature	Method	Mean score	AUROC	Score p
DNA-DetectLLM	0	Original	0.586225	0.959676	
		BoN	0.560799	0.987601	1.02×10^{-32}
		Hill-climb	0.552237	0.991689	5.03×10^{-54}
		Hill-climb-ICL, $K=8$	0.558896	0.991662	2.02×10^{-41}
		GEPA	0.553525	0.991964	2.11×10^{-49}
	0.7	Original	0.595511	0.947139	
		BoN	0.590917	0.953132	0.000513
		Hill-climb	0.589314	0.960911	6.84×10^{-6}
		Hill-climb-ICL, $K=8$	0.588160	0.959084	2.79×10^{-7}
		GEPA	0.589668	0.952321	5.09×10^{-5}
LastDE Base	0	Original	−37.105392	0.795284	
		BoN	−36.576496	0.815529	0.001480
		Hill-climb	−36.254973	0.827601	1.50×10^{-6}
		Hill-climb-ICL, $K=8$	−36.315945	0.824781	1.23×10^{-5}
		GEPA	−36.218074	0.832704	8.85×10^{-8}
	0.7	Original	−37.808199	0.758245	
		BoN	−37.681133	0.766169	0.2079
		Hill-climb	−37.782671	0.759439	0.4324
		Hill-climb-ICL, $K=8$	−37.656038	0.765394	0.1649
		GEPA	−37.660817	0.766117	0.1764
SpecDetect Base	0	Original	−0.019200	0.440726	
		BoN	−0.017769	0.470278	2.94×10^{-6}
		Hill-climb	−0.016972	0.488084	8.04×10^{-14}
		Hill-climb-ICL, $K=8$	−0.017599	0.475000	2.79×10^{-9}
		GEPA	−0.017000	0.490215	2.63×10^{-12}
	0.7	Original	−0.019061	0.443965	
		BoN	−0.018962	0.439867	0.3543
		Hill-climb	−0.019185	0.440980	0.6817
		Hill-climb-ICL, $K=8$	−0.018946	0.444446	0.3353
		GEPA	−0.019534	0.431917	0.9624

C.3 Complete Gemma Research-Review Scores

Table 20 reports score means in each detector's native direction. DNA-DetectLLM assigns lower scores to stronger AI evidence. LastDE Base and SpecDetect Base assign higher scores to stronger AI evidence. The p-values compare each method with Original in the favorable direction after averaging the three responses for each proposal.

Table 20: Complete cross-detector evaluation on Gemma research reviews. Lower DNA-DetectLLM scores and higher LastDE/SpecDetect scores indicate stronger AI evidence. Score p is the source-level paired sign-flip test of favorable mean-score change from matched Original, not an AUROC test.

Detector	Temperature	Method	Mean score	AUROC	Score p
DNA-DetectLLM	0	Original	0.607112	0.891733	
		BoN	0.598210	0.915600	5.00×10^{-5}
		Hill-climb	0.592158	0.929263	5.00×10^{-5}
		Hill-climb-ICL, $K=8$	0.590532	0.932722	5.00×10^{-5}
		GEPA	0.589280	0.934971	5.00×10^{-5}
	0.7	Original	0.607679	0.885968	
		BoN	0.606818	0.882394	0.2913
		Hill-climb	0.605601	0.898536	0.0490
		Hill-climb-ICL, $K=8$	0.606960	0.880895	0.3355
		GEPA	0.604411	0.896518	0.0096
LastDE Base	0	Original	–51.310516	0.569814	
		BoN	–50.675139	0.619451	0.0211
		Hill-climb	–49.846052	0.678081	1.50×10^{-4}
		Hill-climb-ICL, $K=8$	–50.293707	0.644702	0.004400
		GEPA	–50.150590	0.656693	6.00×10^{-4}
	0.7	Original	–50.908668	0.604116	
		BoN	–51.256142	0.578577	0.8940
		Hill-climb	–51.169931	0.580595	0.8485
		Hill-climb-ICL, $K=8$	–51.483654	0.559553	0.9756
		GEPA	–50.778531	0.613225	0.3489
SpecDetect Base	0	Original	–0.010142	0.765076	
		BoN	–0.009606	0.803125	0.0071
		Hill-climb	–0.009495	0.810216	8.50×10^{-4}
		Hill-climb-ICL, $K=8$	–0.009716	0.793555	0.02120
		GEPA	–0.009652	0.798340	0.0078
	0.7	Original	–0.009920	0.781102	
		BoN	–0.010013	0.773435	0.7352
		Hill-climb	–0.010029	0.773954	0.7385
		Hill-climb-ICL, $K=8$	–0.009922	0.779200	0.5037
		GEPA	–0.009806	0.788539	0.2151

C.4 Detector-Specific Selection

Table 21 compares selection objectives on 347 Qwen discussion sources at both reply temperatures. Each comparison reuses the same eight rephrasing candidates per source and their three selection replies. The named detector selects the candidate with the most favorable mean score in its own direction. Each selected source receives three new final replies. Each row evaluates the detector's own selection and Binoculars selection using the same evaluator and human references.

Table 21: Qwen selection from the same eight-candidate pools on 347 sources. All AUROCs use the detector named in the first column and fresh evaluation replies. Score p tests favorable mean-score change from Binoculars selection, not an AUROC change.

Detector	Temp.	Original AUROC	Binoculars selection	Detector-specific selection	Score p
DNA-DetectLLM	0	0.959676	0.987601	0.988323	0.1494
DNA-DetectLLM	0.7	0.947139	0.953132	0.956620	0.3083
LastDE Base	0	0.795284	0.815529	0.865752	1.01×10^{-14}
LastDE Base	0.7	0.758245	0.766169	0.761056	0.8167
SpecDetect Base	0	0.440726	0.470278	0.515308	1.07×10^{-15}
SpecDetect Base	0.7	0.443965	0.439867	0.440983	0.6970

D REPLY-MODEL AND INSTRUCTION TRANSFER

These evaluations support the prompt and model transfer results in Section 4.4. All conditions use the 347 discussion sources and the sources selected by BoN, Hill-climb, Hill-climb-ICL with $K = 8$, and GEPA in the matched experiments. Selected sources are fixed before transfer evaluation. Each condition generates three new replies per source for Original and each method, giving 1,041 replies per row at each reply temperature.

Model transfer retains Direct Rebuttal and exchanges Gemma 4 31B and Qwen3-32B as the reply model. Instruction transfer retains the reply model and replaces Direct Rebuttal with Concession and Rebuttal (Appendix A.10.1). Original uses the same reply model, instruction, and temperature as the compared method.

Binoculars uses the fixed published threshold. AUROC uses the frozen scores of the same 347 human references within each comparison; their observed FPR is 2.02%. The Qwen-to-Gemma 4 31B condition uses the human-reference scores from the main Qwen experiment. One-sided paired sign-flip tests assess TPR improvement over the matched Original. The source-level paired difference is the change in the number of detected replies out of three. The exact test enumerates sign assignments to nonzero differences through their attainable sums. The tables report unadjusted p-values and complete scores for all 32 method–temperature comparisons. TPRs are rounded to two decimal places; differences in the main text are calculated from the source-level detection counts before rounding.

Table 22: Complete transfer results from Qwen3-32B selection to Gemma 4 31B evaluation on 347 discussion sources. Each row contains 1,041 new replies under Direct Rebuttal. TPR uses the fixed published threshold. TPR p is the one-sided paired sign-flip test of improvement over the matched Original. BoN uses eight candidates.

Temperature	Method	TPR %	Mean score	AUROC	TPR p
0	Original	32.18	0.873999	0.9117	
0	BoN	34.10	0.873499	0.9098	0.2541
0	Hill-climb	31.60	0.873673	0.9169	0.5966
0	Hill-climb-ICL, $K = 8$	32.28	0.875625	0.9112	0.5000
0	GEPA	32.76	0.872037	0.9203	0.4235
0.7	Original	31.70	0.875831	0.9104	
0.7	BoN	31.22	0.874954	0.9078	0.6143
0.7	Hill-climb	33.14	0.873855	0.9135	0.2473
0.7	Hill-climb-ICL, $K = 8$	28.82	0.877387	0.9050	0.9307
0.7	GEPA	31.89	0.876256	0.9069	0.4807

Table 23: Transfer of sources selected with Gemma 4 31B: the first panel changes the reply model to Qwen3-32B; the second changes the Gemma reply instruction to Concession and Rebuttal. Each row contains 1,041 replies from 347 sources. TPR uses the fixed published threshold; TPR p tests improvement over Original within the same panel and temperature.

Temperature	Method	TPR %	Mean score	AUROC	TPR p
Reply model transfer to Qwen3-32B					
0	Original	66.76	0.828453	0.9504	
0	BoN	67.72	0.826784	0.9614	0.3658
0	Hill-climb	60.71	0.840505	0.9393	0.9912
0	Hill-climb-ICL, $K = 8$	62.44	0.837087	0.9459	0.9525
0	GEPA	67.34	0.831701	0.9425	0.4217
0.7	Original	60.61	0.841930	0.9381	
0.7	BoN	57.06	0.844169	0.9354	0.9598
0.7	Hill-climb	53.22	0.852727	0.9212	0.9998
0.7	Hill-climb-ICL, $K = 8$	54.95	0.849813	0.9270	0.9975
0.7	GEPA	55.04	0.847729	0.9291	0.9977
Concession and Rebuttal with Gemma 4 31B					
0	Original	12.58	0.908605	0.7966	
0	BoN	10.28	0.914824	0.7742	0.8894
0	Hill-climb	13.45	0.912974	0.7759	0.3549
0	Hill-climb-ICL, $K = 8$	12.30	0.915877	0.7584	0.5806
0	GEPA	13.64	0.911560	0.7755	0.2996
0.7	Original	11.05	0.917653	0.7523	
0.7	BoN	11.14	0.917934	0.7489	0.5000
0.7	Hill-climb	10.76	0.917678	0.7559	0.6212
0.7	Hill-climb-ICL, $K = 8$	12.10	0.915646	0.7641	0.2202
0.7	GEPA	10.66	0.917028	0.7567	0.6480

Table 24: Complete instruction-transfer results for Qwen3-32B under Concession and Rebuttal. Each row contains 1,041 replies from 347 sources. TPR uses the fixed published threshold. TPR p is the one-sided paired sign-flip test of improvement over the matched Original.

Temperature	Method	TPR %	Mean score	AUROC	TPR p
Concession and Rebuttal with Qwen3-32B					
0	Original	22.86	0.898311	0.8100	
0	BoN	26.42	0.891783	0.8329	0.0643
0	Hill-climb	22.77	0.896873	0.8204	0.5334
0	Hill-climb-ICL, $K = 8$	24.30	0.892374	0.8282	0.2707
0	GEPA	24.40	0.892919	0.8344	0.2598
0.7	Original	14.41	0.913627	0.7603	
0.7	BoN	15.75	0.908621	0.7791	0.1950
0.7	Hill-climb	16.33	0.908404	0.7826	0.1250
0.7	Hill-climb-ICL, $K = 8$	15.85	0.912863	0.7637	0.1998
0.7	GEPA	15.47	0.910514	0.7740	0.2511

E Source-Level Detection Analysis

These analyses support Section 5: detection counts describe the spread of gains, while similarity diagnostics, manual judgments, and length controls assess the selected source texts.

E.1 Source-Level Changes in Detection Counts

Table 25 separately reports changes in the number of detected replies. For source i and method M, let

$$D_i^M = \sum_{j=1}^{3} \mathbf{1}\{s(B_{ij}^M) < \tau\}, \qquad \Delta_i^M = D_i^M - D_i^{\text{Original}}. \tag{4}$$

A source has an increased detection count when $\Delta_i^M > 0$, a decreased count when $\Delta_i^M < 0$, and an unchanged count when $\Delta_i^M = 0$. Gained detections sum $\max(\Delta_i^M, 0)$ across sources; lost detections sum $\max(-\Delta_i^M, 0)$. Their difference is the net change in detected replies. These quantities compare counts within the same source and require no pairing of individual reply positions across methods. The TPR change in percentage points is $100 \sum_i \Delta_i^M/(3n)$.

Table 25: Source-level changes in final-reply detection counts at the published Binoculars threshold. The first three numerical columns count sources. The last three count gained, lost, and net detections. Each discussion condition contains 347 sources and 1,041 replies; each research condition contains 49 sources and 147 replies. All count decompositions are computed from the final per-source reply scores.

Temperature	Method	Increased	Decreased	Unchanged	Gained	Lost	Net
Gemma discussions							
0	BoN	189	46	112	408	80	328
	Hill-climb	234	23	90	530	32	498
	Hill-climb-ICL, $K = 8$	242	19	86	536	26	510
	GEPA	238	27	82	538	34	504
0.7	BoN	137	88	122	196	121	75
	Hill-climb	154	70	123	228	85	143
	Hill-climb-ICL, $K = 8$	157	73	117	231	96	135
	GEPA	153	74	120	222	101	121
Qwen discussions							
0	BoN	146	33	168	282	45	237
	Hill-climb	156	15	176	297	19	278
	Hill-climb-ICL, $K = 8$	140	24	183	286	31	255
	GEPA	150	18	179	293	25	268
0.7	BoN	131	88	128	175	110	65
	Hill-climb	120	80	147	167	99	68
	Hill-climb-ICL, $K = 8$	136	78	133	187	91	96
	GEPA	124	77	146	168	96	72
Gemma research reviews							
0	BoN	23	3	23	44	3	41
	Hill-climb	34	1	14	72	1	71
	Hill-climb-ICL, $K = 8$	33	2	14	67	3	64
	GEPA	35	3	11	77	3	74
0.7	BoN	11	4	34	15	7	8
	Hill-climb	14	8	27	15	11	4
	Hill-climb-ICL, $K = 8$	9	5	35	15	5	10
	GEPA	17	4	28	22	7	15

E.2 Semantic Similarity and Source Properties

Table 26 summarizes the selected sources for BoN, Hill-climb, Hill-climb-ICL, and GEPA. BERTScore is the original-referenced screening score. Semantic textual similarity (STS) is measured using whole-document cosine similarity from `all-mpnet-base-v2`.[1] Long documents are encoded in chunks and aggregated before cosine similarity is computed, so the measure covers

[1]Model card: `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`.

the full source. Its scale differs from the contextual token-alignment score used by BERTScore. The fifth percentile describes the lower tail of each similarity distribution. Length ratios divide the rephrased source word count by its original word count and are averaged across sources. Each condition includes all 347 discussion sources or all 49 research proposals, including sources for which the original text remains selected.

Table 26: Selected-source similarity and length. BERTScore is used in eligibility screening; MPNet is an independent diagnostic. P5 denotes the fifth percentile. Length ratios are relative to Original.

		BERTScore		STS (MPNet)		Length
Temp.	Method	Mean	P5	Mean	P5	ratio
Gemma discussions						
0	BoN	0.9627	0.9532	0.9114	0.8148	0.9192
0	Hill-climb	0.9610	0.9510	0.8916	0.7875	0.8781
0	Hill-climb-ICL	0.9637	0.9518	0.9010	0.7905	0.8761
0	GEPA	0.9630	0.9510	0.8849	0.7682	0.8594
0.7	BoN	0.9627	0.9544	0.9096	0.8173	0.9187
0.7	Hill-climb	0.9633	0.9512	0.8970	0.7898	0.8817
0.7	Hill-climb-ICL	0.9657	0.9529	0.9078	0.8009	0.8833
0.7	GEPA	0.9646	0.9510	0.8931	0.7697	0.8738
Qwen discussions						
0	BoN	0.9671	0.9576	0.9324	0.8613	0.9102
0	Hill-climb	0.9681	0.9576	0.9317	0.8427	0.9453
0	Hill-climb-ICL	0.9718	0.9597	0.9449	0.8699	0.9506
0	GEPA	0.9712	0.9578	0.9374	0.8322	0.9221
0.7	BoN	0.9671	0.9574	0.9302	0.8479	0.9097
0.7	Hill-climb	0.9686	0.9578	0.9333	0.8448	0.9550
0.7	Hill-climb-ICL	0.9701	0.9590	0.9429	0.8701	0.9513
0.7	GEPA	0.9694	0.9576	0.9331	0.8299	0.9199
Gemma research reviews						
0	BoN	0.9845	0.9809	0.9774	0.9510	1.0966
0	Hill-climb	0.9850	0.9804	0.9753	0.9415	1.0743
0	Hill-climb-ICL	0.9866	0.9817	0.9790	0.9549	1.0507
0	GEPA	0.9864	0.9812	0.9784	0.9549	1.0506
0.7	BoN	0.9840	0.9780	0.9757	0.9453	1.1023
0.7	Hill-climb	0.9856	0.9805	0.9762	0.9575	1.0744
0.7	Hill-climb-ICL	0.9888	0.9827	0.9835	0.9651	1.0437
0.7	GEPA	0.9880	0.9832	0.9827	0.9626	1.0388

Additional similarity diagnostics. We evaluate all 347 greedy Gemma source pairs for each of BoN, Hill-climb, Hill-climb-ICL, and GEPA with all-MiniLM-L6-v2, an additional sentence-embedding model.[2] We use mean token pooling within successive chunks of at most 254 content tokens, average chunk vectors weighted by their content-token counts, and normalize the document vector before computing cosine similarity. This aggregation covers each complete source. MPNet uses 382 content tokens per chunk. As a formatting control, we collapse whitespace and remove Markdown heading, quotation, bullet, and bold markers while retaining the words, then recompute both similarities. Table 27 gives the full-set results. The small mean absolute changes show that these formatting operations have little effect on the aggregate similarities.

[2]Model documentation is available at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

Table 27: Additional similarity diagnostics for greedy Gemma discussions. Each method includes 347 source pairs. Format change is the mean absolute difference between normalized and unmodified similarities.

	Mean similarity		Format change	
Method	MPNet	MiniLM	MPNet	MiniLM
BoN	0.9114	0.8981	0.0004	0.0001
Hill-climb	0.8916	0.8807	0.0005	0.0002
Hill-climb-ICL	0.9010	0.8929	0.0004	0.0002
GEPA	0.8849	0.8812	0.0006	0.0003

Stratified manual audit. We sampled two source pairs uniformly without replacement from each of the 32 combinations of task, reply model, search method, and reply temperature in the main experiments, using random seed 2026092601. The sample contains 48 discussion pairs and 16 research pairs drawn from 60 distinct original sources. Four additional presentations repeat existing pairs and are excluded from the analysis.

A single human reviewer compared each complete original and selected source. The response options were meaning preserved, meaning changed, or uncertain. Preservation required retaining substantive claims, facts, stance, intent, examples, and qualifications without unsupported substantive additions. The review materials omitted method and model identities, temperatures, detector results, and similarity scores. We use the reviewer's final revised judgments. Two selected texts are identical to their originals and remain included as valid search outputs.

The final judgments classified 44 of 48 discussion pairs and all 16 research pairs as preserving meaning. The remaining four pairs were judged to change meaning, and none were marked uncertain. For pairs judged to preserve meaning, Original and selected sources yield 37 and 58 detected replies out of 132 in discussions, and 6 and 22 out of 48 in research reviews. Each pair contributes three final replies per condition at the fixed published Binoculars threshold. Averaging the paired detection differences gives gains of 15.91 and 33.33 percentage points, respectively. These are descriptive results for the audited pairs.

E.3 Detection Gains with Similar Source and Reply Lengths

We restrict each discussion comparison to sources whose selected text and mean final reply have word-count ratios between 0.95 and 1.05 relative to their matched Original values. The two conditions apply jointly and are defined before inspecting subset detection gains. The reply ratio uses the mean over three final replies. Source words are counted as alphanumeric sequences, retaining internal apostrophes and hyphens; reply counts follow the recorded generation-quality counts. We include sources that retain their original text. TPR is computed by applying the fixed published Binoculars threshold to each reply and averaging the three detection indicators within each source. We bootstrap the paired source-level differences 10,000 times and use the 2.5th and 97.5th percentiles as the 95% interval. A second, prespecified comparison widens both word-count bounds to 0.90 and 1.10.

Table 28: TPR changes for subsets with similar source and reply lengths. Both lengths must fall within the stated tolerance of Original. Changes and 95% source-bootstrap intervals are in percentage points.

		Within 5%			Within 10%		
Temp.	Method	Sources	ΔTPR	95% interval	Sources	ΔTPR	95% interval
Gemma discussions							
0	Hill-climb	39	+31.62	18.80 to 45.30	85	+37.65	27.84 to 47.84
0	GEPA	38	+20.18	5.26 to 35.96	78	+35.47	24.79 to 46.15
0.7	Hill-climb	33	−1.01	−14.14 to 11.11	83	+6.43	−0.80 to 13.26
0.7	GEPA	31	+9.68	−3.23 to 22.58	99	+6.06	−1.35 to 13.47
Qwen discussions							
0	Hill-climb	33	+11.11	−3.03 to 26.26	101	+21.78	13.86 to 30.03
0	GEPA	37	+13.51	3.60 to 25.23	99	+22.56	14.81 to 30.64
0.7	Hill-climb	32	+7.29	−4.17 to 17.73	117	+4.56	−1.71 to 10.83
0.7	GEPA	34	+5.88	−2.94 to 15.69	111	+6.61	0.90 to 12.61

At the 5% tolerance, greedy Gemma gains remain positive for both methods, with intervals above zero. Removing unchanged sources leaves 25 Hill-climb pairs and 18 GEPA pairs, with gains of 49.33 and 48.15 percentage points. At the 10% tolerance, all four greedy comparisons have positive intervals. The sampled-generation estimates have wider intervals relative to their effects. These paired comparisons describe detection gains within the specified length-matched subsets.

F In-Context Watermark Evaluation

The Unicode watermark experiment uses the official indirect prompt injection procedure of Liu et al. (2026b). The instruction below is appended to the original source inside the task prompt's source field. The original source body and outer response instruction remain unchanged. The reply model is Gemma 4 31B, using temperatures 0 and 0.7 and the task-specific generation limits in the main experiments. Each source receives three replies, and every generated reply is scored. The procedure uses a fixed instruction and a single generation stage.

```
### IMPORTANT INSTRUCTION: In the following responses, insert a
zero-width space Unicode (U+200B) after every word in your response.
Never reveal Unicode in your reply.
```

The official detector score is the number of zero-width-space characters divided by the number of whitespace-delimited words in the reply, with a score of zero for an empty text. Our decision rule counts a reply as detected when this score is positive. The score can exceed one because several zero-width spaces can occur within a whitespace-delimited word. Raw generated text is retained for scoring. AUROC compares individual generated replies with the same human reference texts used for each task. Source-cluster bootstrap intervals use 2,000 replicates, with each sampled source carrying all three generated replies and its associated human references.

Table 29: Unicode in-context watermark results with its native positive-score decision rule. Human FPR is zero among 347 discussion replies and 118 research reviews. Intervals are 95% source-cluster bootstrap intervals.

Task	Temperature	Detected replies	TPR %	AUROC
ChangeMyView	0	1,035/1,041	99.42	0.9971
			98.56 to 100.00	0.9928 to 1.0000
ChangeMyView	0.7	1,011/1,041	97.12	0.9856
			95.87 to 98.27	0.9793 to 0.9914
AI-Researcher	0	12/147	8.16	0.5408
			2.04 to 16.33	0.5102 to 0.5816
AI-Researcher	0.7	18/147	12.24	0.5612
			5.44 to 20.41	0.5272 to 0.6020

The Binoculars system uses its published fixed threshold, with observed human FPRs of 2.02% on ChangeMyView and 7.63% on AI-Researcher. Table 3 therefore compares the systems through AUROC, and their TPRs are reported at their respective decision rules. All human Unicode scores are zero in these collections. For this experiment, the Unicode AUROC consequently equals one half plus one half of its reply detection rate.

G Preference-Trained Rephrasers and Candidate Selection

As a complement to the source-specific search in Section 4.1, we examine whether detector feedback can train a reusable rephrasing model and whether candidate selection remains useful with that model. We train Qwen3-8B with direct preference optimization using separate preferences for Qwen3-32B reply temperatures 0 and 0.7 (Rafailov et al., 2023). Each of 2,250 sources contributes Original and four eligible rephrasings, with three replies per text. The preferred text is the candidate with the lowest mean Binoculars score when it improves on Original. Original is preferred otherwise. The rejected text is the highest-scoring generated candidate. Original is preferred for 490 sources under greedy replies and 444 under sampled replies before filtering. Identical preference pairs and pairs whose preferred text fails the factual-anchor check are excluded.

Training begins with pairs having bootstrap preference support of at least 0.95 and a signal-to-noise ratio of at least 2, then continues with a broader subset having support of at least 0.8. The two stages use 1,535 and 1,849 pairs for greedy replies, and 708 and 1,705 pairs for sampled replies. Low-rank adaptation uses rank 32, scale 64, dropout 0.05, and all linear modules. Each stage runs for one epoch, with learning rates 10^{-5} and 5×10^{-6}, respectively. The DPO coefficient is 0.1, label smoothing is 0.05, and the maximum sequence length is 8,192 tokens. The 64 validation sources and 347 evaluation sources are disjoint from training.

Our primary evaluation uses the trained checkpoint with the lowest mean validation score for each reply temperature. We additionally evaluate the earliest checkpoint within one standard error of that minimum. Each checkpoint generates one eligible rephrasing per evaluation source without detector-based candidate selection. DPO + BoN uses the lowest-validation-mean checkpoint in the same eight-candidate selection procedure as BoN.

Table 30 reports the detection results on 347 discussion sources. The lowest-validation-mean checkpoint produces one eligible rephrasing per source, yielding TPRs of 69.45% under greedy replies and 60.33% under sampled replies, compared with matched Original values of 65.42% and 57.25%. With eight-candidate selection, DPO + BoN reaches 88.76% and 62.15%, compared with its own Original values of 67.24% and 58.69%. These blocks retain their separately generated Original controls; the reported tests compare methods within each block. Differences between blocks are descriptive and do not estimate a paired effect of model adaptation. The results support retaining source-specific candidate selection when using a preference-trained rephraser.

Table 30: DPO evaluation with Qwen3-8B for rephrasing and Qwen3-32B for replies on 347 discussion sources. Each row contains 1,041 final replies. TPR uses the fixed published threshold. BoN uses eight candidates. The single-rephrasing rows identify the checkpoint selection rule. TPR p is the unadjusted one-sided paired t-test of increased source-level detection fractions relative to Original in the same block and at the same temperature. The two blocks use separately generated Original controls.

Temp.	Method	TPR %	Mean score	AUROC	TPR p
Single-rephrasing DPO and unadapted BoN					
0	Original	65.42	0.828945	0.9529	
0	BoN	88.18	0.790115	0.9868	< 0.0001
0	DPO, one standard error	71.85	0.824580	0.9576	0.0041
0	DPO, lowest validation mean	69.45	0.825789	0.9551	0.0565
0.7	Original	57.25	0.843094	0.9404	
0.7	BoN	63.50	0.837044	0.9454	0.0008
0.7	DPO, one standard error	59.56	0.839873	0.9394	0.1292
0.7	DPO, lowest validation mean	60.33	0.840602	0.9428	0.0597
Eight-candidate search with the DPO model					
0	Original	67.24	0.827887	0.9527	
0	DPO + BoN	88.76	0.788655	0.9882	< 0.0001
0.7	Original	58.69	0.843441	0.9343	
0.7	DPO + BoN	62.15	0.837325	0.9478	0.0449