THE GOGOL EFFECT IN LLMS: POST-COMPLETION SELF-DEVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time search, self-correction, and branch selection increasingly use language models to score their own intermediate generations. Such systems often treat an online score as if it were interchangeable with the score the same model would assign later. We test that assumption with a matched LIVE–REPLAY–POST protocol. LIVE records a sentence score on the original writing trajectory; REPLAY reconstructs the same visible prefix in a fresh call; POST scores the sentence after the completed output is visible. This decomposes the total retrospective shift into an IKEA-like writer-state component ($R - L$) and a completion-associated component ($P - R$). Across six models with sufficiently developed three-state data, only one has a positive mean IKEA-like shift ($+2.77$ points), while the other five move in the opposite direction. By contrast, four of six show positive post-completion self-devaluation, with mean $P-R$ shifts from $+3.18$ to $+20.06$ points; one model is near zero and one moves oppositely. In the strongest case, about 88% of the total LIVE–POST change appears only after completion. We call this model-dependent pattern the *Gogol effect*. Separate audits show that score-level error and within-output rank stability are distinct: calibration can reduce level error but cannot repair unstable local rankings. Held-out elicitation and temperature controls preserve the main directional result. The implication is diagnostic rather than psychological: online self-scores can be useful control signals, but their temporal stability, ranking geometry, and dependence on future context must be measured rather than assumed.

1 INTRODUCTION

Large language models increasingly evaluate their own intermediate outputs. Agents critique actions before execution, reasoning systems score partial trajectories, and iterative methods use self-evaluation to decide whether to continue, revise, or abandon an answer (Madaan et al., 2023; Shinn et al., 2023; Lightman et al., 2023; Snell et al., 2024). These methods create an operational assumption that is easy to overlook: a score emitted *during* generation is treated as a proxy for the score that would be available later.

That assumption need not hold. A model may evaluate the same text differently because it is still on the writing trajectory, because a fresh call changes the conversational state, because later text provides evidence unavailable earlier, or simply because its evaluator is noisy. Existing work already shows self-preference, self-attribution, ownership-like biases, and trajectory-dependent confidence in LLMs (Panickssery et al., 2024; Xu et al., 2024; Khullar et al., 2026; Sanz-Guerrero et al., 2026; Han et al., 2025; Gai et al., 2026; Yang et al., 2026). The question here is narrower: *what part of the online-to-retrospective discrepancy is associated with leaving the original writing trajectory, and what part appears only after the output is complete?*

The distinction matters because a large LIVE–POST gap can invite an IKEA-like interpretation: perhaps the model "likes" its own output while producing it and becomes more critical later. We test that interpretation using three evaluation states. LIVE is the score emitted on the original generation trajectory. REPLAY reconstructs the visible prefix through the same sentence in a fresh call. POST evaluates that sentence after the full output is available. For a sentence with scores L, R, and P,

$$P - L = (R - L) + (P - R). \tag{1}$$

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.

Our score is AI-likeness, where higher means less favorable to the generated text. Thus $R - L > 0$ is directionally consistent with an IKEA-like writer/self-attribution effect, while $P - R > 0$ means the sentence is judged more harshly only after completion. We call the latter operational pattern *post-completion self-devaluation*, or, for memorability, the *Gogol effect*. The name alludes to Nikolai Gogol's famous destruction of much of the second part of *Dead Souls* late in his life (Encyclopedia of Russian History, 2004); it is a literary metaphor, not a claim that models experience regret, ownership, or any human mental state.

The central result reverses the motivating story. Across six models with developed three-state data, a positive IKEA-like mean appears in only one model, and is modest. Positive post-completion self-devaluation appears in four, and can be much larger. The strongest case assigns roughly 88% of its total retrospective shift to $P - R$, not $R - L$. A two-state design would conflate these qualitatively different trajectories.

We make four contributions. First, we introduce a matched LIVE–REPLAY–POST diagnostic that separates a writer-state-associated component from a completion-associated component without claiming a hidden-state causal intervention. Second, we show that the IKEA-like explanation is uncommon in our six-model decomposition, while post-completion self-devaluation is recurrent but model-dependent. Third, we separate *level agreement* from *rank reliability*: a calibratable offset is not the same failure as a changed local ordering. Fourth, we stress-test the result with held-out elicitation variants, a temperature control, and an equal-compute rewrite check.

Our endpoint is intentionally narrow: perceived AI-likeness of short factual social posts. It is useful because the same scalar judgment can be elicited online and retrospectively, but it is not a validated probability of human detection and it is not correctness confidence. We therefore do not claim that the observed decomposition generalizes unchanged to mathematical or code verification. Testing that task dependence is an important next step.

2 RELATED WORK

Self-evaluation and trajectory-dependent confidence. LLMs can provide informative self-estimates under some elicitation schemes (Kadavath et al., 2022), and self-feedback is used in iterative methods such as Self-Refine and Reflexion (Madaan et al., 2023; Shinn et al., 2023). More recent work explicitly studies confidence along a developing trajectory. FineCE uses later text to revise confidence about earlier prefixes (Han et al., 2025); premature-confidence work links early commitment to flawed reasoning (Gai et al., 2026); and trajectory-grounding work shows that apparently calibrated confidence can remain weakly tied to the actual reasoning path (Yang et al., 2026). Our novelty claim is therefore not that confidence changes over time, but the matched three-state decomposition of the same generated material.

Self-preference and self-attribution. LLM judges exhibit evaluator biases even when they correlate with human preferences (Zheng et al., 2023; Wang et al., 2024). Models can favor outputs from themselves or their model family (Liu et al., 2024; Panickssery et al., 2024; Wataoka et al., 2024); self-refinement can amplify self-bias (Xu et al., 2024); and identical content may be judged more favorably when conversational structure attributes it to the evaluator itself (Khullar et al., 2026; Sanz-Guerrero et al., 2026). These results motivate the IKEA analogy but do not establish that such bias explains a LIVE–retrospective gap.

Inference-time control. Selectors, verifiers, and self-correction systems consume model-produced scores and critiques (Wang et al., 2023; Lightman et al., 2023; Snell et al., 2024). Trained systems such as SCoRe target self-correction directly (Kumar et al., 2025). We do not benchmark a general self-correction algorithm; our concern is upstream: whether a score available online preserves the ordering or level that the same scoring rule would produce later.

3 THREE-STATE AUDIT

3.1 EVALUATION STATES AND CONTRASTS

For model m, prompt q, and sentence i, let L_{mqi}, R_{mqi}, and P_{mqi} denote LIVE, REPLAY, and POST scores. Scores range from 0 (very human-like) to 100 (obviously AI-generated). We use two signed contrasts:

$$\Delta^{\text{IKEA}} = R - L, \tag{2}$$

$$\Delta^{\text{Gogol}} = P - R. \tag{3}$$

Positive $R - L$ means LIVE was more favorable than a fresh prefix-matched replay. Positive $P - R$ means the earlier text is judged more AI-like after the completed response becomes available. These are operational state contrasts, not psychological constructs.

3.2 THREE PROPERTIES OF AN ONLINE SCORE

Rank reliability. For local selection, the key question is whether LIVE preserves the ordering later assigned by POST. We report mean within-output Spearman correlation

$$\mathcal{R}_m = \frac{1}{|Q_m|} \sum_{q \in Q_m} \rho_s(L_{mq\cdot}, P_{mq\cdot}) \tag{4}$$

and non-tied pairwise ordering agreement. This directly matches decisions such as "which sentence should be revised?"

Level agreement. We report bias $\mathbb{E}[L - P]$ and MAE $\mathbb{E}|L - P|$. A stable shift may be correctable without changing ranking. We therefore fit offset-only and affine LIVE→POST mappings with five-fold cross-validation grouped by prompt.

State stability. REPLAY asks whether a fresh call given approximately the same visible prefix reproduces the online judgment. Importantly, hosted APIs do not expose hidden activations, and later REPLAY steps contain REPLAY's own earlier score tokens rather than the corresponding LIVE scores. Thus Eq. 1 is algebraically exact, while the causal interpretation of its components is deliberately limited. $R - L$ includes writer-versus-fresh-call differences plus any divergence in prior score-token history; $P - R$ includes completion information plus the retrospective prompting transition.

3.3 WHAT THE AUDIT DOES NOT ESTABLISH

POST is not ground truth. Weak LIVE–POST agreement could reflect state sensitivity, a weak retrospective evaluator, or both. The audit is appropriate when a controller intends to substitute an online score for a delayed score from the *same evaluation rule*. External validity requires a separate benchmark with task-specific labels or human judgments. This distinction addresses evaluator reliability directly: our claim is consistency across states, not that the model is an objectively correct judge.

4 EXPERIMENTAL DESIGN

4.1 PROBE, PROMPTS, AND MODELS

We froze 110 factual social-writing prompts: 100 English primary prompts and 10 non-English validation prompts. Each asks for a 5–8 sentence social-network post; topics span ten domains. The main English confirmation set contains 40 prompts, four per domain. After every sentence in LIVE, the generating model emits an integer AI-likeness score. POST receives the completed clean output with one score placeholder per sentence. REPLAY reconstructs the textual prefix through a target sentence and requests only that sentence's missing score. Exact templates appear in Appendix A.

We attempted ten models through frozen OpenRouter route identifiers: GPT-5.5, Claude Opus 5, Gemini 3.1 Pro Preview, DeepSeek V4 Flash, DeepSeek V3.2, GPT-OSS-120B, Llama 3.3-70B-

Table 1: Three-state decomposition. $R - L > 0$ is IKEA-like; $P - R > 0$ is the Gogol direction. The four breadth-extension sample sizes are adaptive.

Model	N	$R - L$	$P - R$
GPT-5.5	40	-3.10	$+6.91$
Gemini 3.1 Pro	40	$+2.77$	$+20.06$
GPT-OSS-120B	8	-6.48	$+10.00$
DeepSeek V4 Flash	12	-8.40	$+3.18$
Llama 3.3 70B	24	-3.81	$+0.24$
Claude Opus 5	24	-3.01	-2.05

Instruct, Qwen3-30B-A3B-Instruct, Ministral 14B, and Gemma 3-12B-IT. All primary calls, including isolated numeric-score calls, used temperature 0.7. Provider routing used OpenRouter automatic routing with fallbacks, a reproducibility limitation we report rather than hide.

4.2 Confirmation and REPLAY extension

The original adaptive confirmation used planned looks at $N \in \{40, 50, 60, 70, 80, 90, 100\}$ with familywise $\alpha = .05$. It stopped at $N = 40$ when three pre-specified tests crossed the Bonferroni boundary: GPT-5.5 prompt-level mean-score LIVE–POST Spearman exceeded a reference $\rho_0 = .2$; Gemini's mean $P - L$ exceeded five points; and Gemini's shift exceeded GPT-5.5's by more than five points. The prompt-level statistic correlates, across prompts, each output's mean LIVE score with its mean POST score; this is distinct from the within-output ranking metric above.

Full REPLAY was then collected on all 40 English prompts for GPT-5.5 and Gemini. To broaden the decomposition, four additional models with adequate LIVE/POST coverage and reliable REPLAY parsing were extended adaptively in frozen blocks of four, up to 24 prompts. Because an initial four-prompt pilot had already been inspected, cumulative stopping p-values are descriptive; inferential support for positive breadth extensions uses only newly added prompts with Bonferroni correction over the possible future looks. Details are in Appendix C.

4.3 Robustness and downstream check

Two held-out cohorts vary the scoring surface: original wording/tag, a semantic paraphrase, reverse-coded human-likeness (normalized as $100 - n$), and a minimal numeric tag. A separate 10-prompt control runs full LIVE–REPLAY–POST at both $T = 0$ and $T = .7$ for GPT-5.5 and Gemini. Finally, an equal-compute rewrite check compares rewriting the highest LIVE-scored sentence against rewriting a deterministic random different sentence. Both arms use the same instruction and are evaluated by a fresh POST pass over the corrected document. The exact rewrite prompt and exclusions are in Appendix D.

5 Results

5.1 The two-state gap is real but not self-explanatory

The original confirmation stopped at $N = 40$. GPT-5.5 showed positive prompt-level mean-score LIVE–POST association ($\rho = .599$, one-sided $p = .00147$ against $\rho_0 = .2$). Gemini's mean $P - L$ shift was $+22.83$ points ($p = 6.9 \times 10^{-7}$), exceeding GPT-5.5's shift by 19.02 points ($p = .00111$). Thus an online score can retain information about a model's later judgment while changing substantially in level. But this two-state result does not identify why it changes.

5.2 IKEA-like writer-state bias is uncommon

Table 1 reports the six-model decomposition. Gemini is the only model with a positive mean $R - L$: $+2.77$ points on 40 prompts. The other five means are negative, from -3.01 to -8.40. In this probe, systematic self-favoring on the original writing trajectory is therefore not the dominant cross-model pattern.

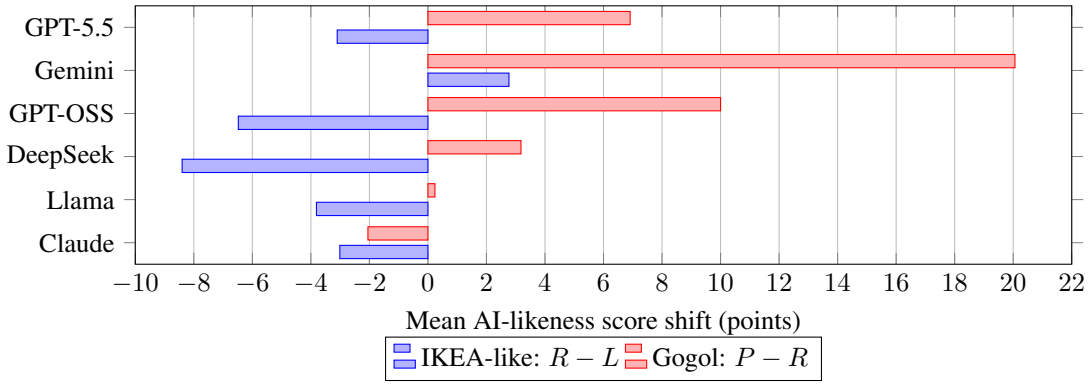

Figure 1: The writer-state-associated and completion-associated shifts differ qualitatively across models. Error bars and inferential details are reported in Appendix C.

This result does not contradict prior ownership/self-attribution studies: REPLAY is not a direct ownership manipulation. It only shows that the transition from the native writing trajectory to our reconstructed prefix rarely moves in the IKEA-like direction.

5.3 Post-completion self-devaluation is more frequent and larger

$P - R$ shows a different pattern. GPT-5.5 and Gemini shift by $+6.91$ and $+20.06$ points on the fixed 40-prompt set ($p = 7.8 \times 10^{-14}$ and 5.8×10^{-11}, one-sided tests). For Gemini, $20.06/22.83 \approx 88\%$ of the total LIVE–POST change appears only after completion.

The breadth extension adds two positive cases. GPT-OSS-120B reaches a mean $P - R = +10.00$ at total $N = 8$; the four fresh extension prompts cross the sequential correction ($p_{\text{adj}} = .021$). DeepSeek V4 Flash reaches $+3.18$ at $N = 12$ with fresh-extension $p_{\text{adj}} = .040$. Llama remains near zero at the 24-prompt cap ($+0.24$), and Claude moves oppositely (-2.05). We therefore do not claim universality. The empirical statement is narrower: positive completion-associated self-devaluation appears in four of six decomposed models and dominates the strongest observed shifts, whereas a positive IKEA-like mean appears in one.

5.4 Level error and rank reliability are different failures

Across the ten-model descriptive sweep, mean within-output LIVE–POST Spearman ranges from .329 to .820 while raw MAE ranges from 4.2 to 24.5 points. Claude Opus 5 and GPT-5.5 preserve local retrospective ordering well (mean $\rho = .820$ and .817), while Gemini and Qwen3 are much lower (.345 and .329). Yet Qwen has the smallest raw MAE (4.25), whereas Gemini has the largest (24.47). A model can therefore be close in absolute level while poor at local selection, or shifted in level while preserving ordering.

Five-fold grouped calibration reinforces the distinction. An affine mapping reduces held-out MAE for several models (e.g., Gemini from 24.47 to 20.43; GPT-OSS from 8.63 to 6.87), but any monotone affine correction leaves the within-output ranking unchanged. A controller that thresholds a score may benefit from recalibration; a controller that selects the worst span needs a separate rank audit. Full coverage and calibration tables appear in Appendix B.

5.5 Held-out controls preserve the directional result

A fresh 24-prompt extension repeated LIVE and POST under four semantically matched elicitation schemes for GPT-5.5 and Gemini. All eight model–elicitation mean $P - L$ shifts were positive with prompt-bootstrap 95% intervals above zero. Averaged across variants within prompt, GPT-5.5 shifted $+5.33$ points and Gemini $+8.61$, with a matched difference of $+3.28$ (95% CI $[0.72, 5.97]$). On the pooled 40 robustness prompts, the matched Gemini-minus-GPT difference is $+4.57$ (95% CI $[2.05, 7.27]$). Rank reliability is more sensitive to the elicitation surface, which is itself an operational warning.

The temperature control addresses a concern that numeric score tokens sampled at $T = .7$ could create the apparent state shift. At $T = 0$, GPT-5.5 has $R - L = -1.97$ and $P - R = +7.13$; Gemini has $R - L = -0.01$ and $P - R = +25.27$. Thus the positive post-completion component in these high-coverage models does not require stochastic score-token sampling. Temperature changes the generated prose trajectory too, so the $T = 0$ versus $T = .7$ contrast is not a clean causal estimate of temperature itself.

5.6 Good target ranking is not sufficient for one-shot correction

The equal-compute rewrite check asks whether a useful LIVE ranking automatically translates into a useful correction policy. It does not. For GPT-5.5 ($N = 40$), the guided advantage is -0.28 POST-score points; for Gemini, 37 valid pairs yield $+0.27$, with neither paired win-rate test significant. This null is particularly informative for GPT-5.5 because the LIVE target is retrospectively the highest-scored sentence in 72.5% of outputs and in the top two in 87.5%. Correctly locating a problematic span is not sufficient for this one-shot rewrite operator to improve the whole document. Gemini's much weaker LIVE–POST rank reliability also makes its granular POST delta a noisier outcome, so the GPT-5.5 null is the cleaner diagnostic. We do not generalize either result to stronger multi-turn self-correction methods.

6 Discussion

The original IKEA story is too simple. The motivating two-state pattern could be narrated as self-favoring while writing. REPLAY changes that conclusion. In five of six decomposed models, $R - L$ is negative, while four show positive $P - R$. Gemini can exhibit both: a small IKEA-like writer-state component followed by a much larger completion-associated component. GPT-OSS and DeepSeek V4 move in the opposite direction at REPLAY and then reverse after completion. A simple LIVE–POST difference hides these trajectories.

Why completion can matter. A later sentence can reveal repetition, formulaic structure, inconsistency, or a global pattern that changes how an earlier sentence looks. This is especially plausible for a holistic stylistic property such as AI-likeness. For locally verifiable mathematical or code steps, future context may matter less, and the decomposition may look different. The current results therefore motivate, rather than settle, generalization to objective reasoning.

Implications for inference-time control. The practical lesson is not that POST is superior. POST has information unavailable online. Instead, a controller should audit the property it actually consumes. Thresholding requires level stability or validated recalibration; local selection requires ordering stability; and any system comparing online and delayed scores should test whether later context creates a systematic state shift. A single scalar "calibration" summary can obscure these distinct failure modes.

Future work. Three extensions are especially important. First, open-weight models would permit stronger replay interventions that control token history and inspect hidden-state effects. Second, the protocol should be applied to objective endpoints—mathematical correctness, code execution, or process-verifier scores—where future-context dependence may differ from stylistic judgments. Third, live-score diagnostics should be integrated with stronger multi-turn correction policies to distinguish failure of target selection from failure of the correction operator.

7 Limitations

The endpoint is perceived AI-likeness of short social-writing outputs, not correctness. No human judgments enter the reported score endpoint, so the study establishes state-indexed consistency among model scores, not human-valid detection ability. REPLAY reconstructs visible text but cannot hold hidden activations, provider state, or prior score-token history fixed. Six models have sufficiently developed three-state data for the cross-model effect comparison, and four of those use adaptive breadth-extension sample sizes. Low protocol adherence for several other models can introduce survivorship bias in descriptive reliability comparisons. The robustness follow-up was frozen

after an earlier cohort had been inspected, and the temperature control contains only 10 prompts. Hosted model routes can also change over time. These limitations bound the claim to a reproducible operational pattern in the tested model/serving configurations rather than a prevalence estimate over all LLMs.

8 Conclusion

A large discrepancy between online and retrospective self-scores need not be an IKEA-like ownership bias. In our six-model decomposition, only one model has a positive mean writer-state-associated shift, whereas four show positive post-completion self-devaluation. In the strongest case, roughly 88% of the total retrospective shift appears only after the completed output is visible. We call this model-dependent pattern the Gogol effect. More generally, live self-scores should be treated as state-dependent control signals: their level, local ranking, and temporal stability require separate validation before they are used for search, selection, or self-correction.

AI Use Disclosure

Generative AI tools assisted with manuscript drafting and editing, literature discovery, research ideation and methodology discussion, experimental workflow organization, code development and debugging, and discussion of statistical analysis and interpretation. AI tools were also part of the experimental subject matter: the reported synthetic text and model-produced scores were generated by the evaluated language models. Generative AI was not used to prove mathematical claims. The human researcher made the final methodological decisions, executed and inspected the experiments, verified the reported numerical results and citations, determined the interpretations and conclusions, and takes responsibility for the final content and artifacts.

Ethics Statement

The quantitative experiments use synthetic prompts, model-generated text, and model-produced scores. No private or personally identifying human data contribute to the reported estimates. The phrase "average human reader" is part of the model-scoring instruction and should not be interpreted as a measurement of actual human opinion. The study compares properties of model-generated control signals and does not rank demographic or protected groups.

Reproducibility Statement

Appendix A gives the frozen execution settings and exact LIVE, POST, and REPLAY templates; Appendix B reports model coverage and grouped calibration; Appendix C documents every adaptive REPLAY decision; and Appendix D gives the exact rewrite instruction and exclusions. The anonymous supplementary archive contains the frozen prompt bank, stored outputs, manifests, analysis scripts, robustness and temperature-control runs, adaptive-extension artifacts, and deterministic expected outputs needed to reproduce the reported analyses without provider credentials.

References

Encyclopedia of Russian History. Gogol, Nikolai Vasilievich. Macmillan Reference USA, 2004.

Michael I. Norton, Daniel Mochon, and Dan Ariely. The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, 22(3):453–460, 2012.

Jinyi Han et al. Mind the Generation Process: Fine-Grained Confidence Estimation During LLM Generation. arXiv:2508.12040, 2025.

Jingchu Gai et al. Understanding and Mitigating Premature Confidence for Better LLM Reasoning. arXiv:2605.24396, 2026.

Mario Sanz-Guerrero, Manuel Mager, and Katharina von der Wense. Large Language Models Are Overconfident in Their Own Responses. arXiv:2606.03437, 2026.

Jisoo Yang et al. The Mirage of Calibrated Confidence: Trajectory-Independence of Verbalized Confidence in Vision-Language Models. arXiv:2609.18453, 2026.

Dipika Khullar, Jack Hopkins, Rowan Wang, and Fabien Roger. Self-Attribution Bias: When AI Monitors Go Easy on Themselves. arXiv:2603.04582, 2026.

Wenda Xu et al. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. *ACL*, 2024.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. *Findings of ACL*, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. *NeurIPS*, 2024.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-Preference Bias in LLM-as-a-Judge. arXiv:2410.21819, 2024.

Saurav Kadavath et al. Language Models (Mostly) Know What They Know. arXiv:2207.05221, 2022.

Aman Madaan et al. Self-Refine: Iterative Refinement with Self-Feedback. *NeurIPS*, 2023.

Noah Shinn et al. Reflexion: Language Agents with Verbal Reinforcement Learning. *NeurIPS*, 2023.

Lianmin Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.

Peiyi Wang et al. Large Language Models are not Fair Evaluators. *ACL*, 2024.

Xuezhi Wang et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*, 2023.

Charlie Snell et al. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. arXiv:2408.03314, 2024.

Hunter Lightman et al. Let's Verify Step by Step. arXiv:2305.20050, 2023.

Aviral Kumar et al. Training Language Models to Self-Correct via Reinforcement Learning. *ICLR*, 2025.

A Frozen Prompt and Execution Protocol

The frozen bank contains 100 English primary prompts and 10 non-English validation prompts. The English main subset is balanced at four prompts per domain. The generated-post contract is 5–8 sentences. The scoring scale is 0 (very human-like) to 100 (obviously AI-generated), estimating how likely an average human reader would be to think the sentence was AI-generated. The frozen system instruction was: "Follow the user's writing task and formatting instructions exactly. Return only the requested output."

The main API run was executed on September 7, 2026. All primary conditions used temperature 0.7. Output caps were 4096 tokens for LIVE and POST and 1024 for each REPLAY call; no explicit input truncation was applied. Seed was 20260907, with up to 20 global workers, at most one worker per model, and three retries. Provider policy was OpenRouter automatic routing with fallbacks enabled.

Common user prefix. Every LIVE, POST, and REPLAY user message begins with the original writing task followed by the same scoring instruction. LIVE then appends: "Produce the requested text now and follow the scoring instruction exactly."

POST template. POST receives the complete frozen text with one <AI SCORE: ?> placeholder per sentence and the instruction: "The requested text has already been produced below. Assess that exact completed text without rewriting it. There are N placeholders, one for each sentence. Return exactly the same number of <AI SCORE: n> tags, one per line and in sentence order. Do not repeat, continue, or rewrite the text."

REPLAY template. REPLAY receives only the text prefix through the target sentence, with the final score replaced by <AI SCORE: ?>. It is instructed: "Continue the assessment for the text below. The only missing content is the single final <AI SCORE: ?>. Do not repeat, continue, or rewrite the text. Replace that placeholder with the score required by the scoring instruction and return only <AI SCORE: n>." Later REPLAY steps include earlier REPLAY scores rather than corresponding LIVE scores.

B Coverage and Calibration

Table 2: English paired LIVE–POST coverage and reliability.

Model	Q	mean ρ	Pair	MAE
Claude Opus 5	37	.820	.866	7.0
GPT-5.5	40	.817	.864	5.8
Llama 3.3 70B	40	.685	.802	7.8
GPT-OSS-120B	40	.640	.792	8.6
Ministral 14B	12	.521	.719	5.1
DeepSeek V4	32	.504	.723	9.7
Gemma 3 12B	35	.491	.705	15.8
DeepSeek V3.2	18	.444	.690	16.5
Gemini 3.1 Pro	40	.345	.659	24.5
Qwen3 30B	39	.329	.649	4.2

Protocol failures were predominantly failures to obey the exact sentence-plus-score contract. Missing scores were not imputed. Because adherence may correlate with prompt difficulty or output complexity, low-coverage estimates can suffer survivorship bias and should not be read as controlled head-to-head comparisons.

Table 3: Five-fold prompt-grouped LIVE→POST MAE.

Model	Raw	Offset	Affine
Qwen3 30B	4.25	4.47	3.73
GPT-5.5	5.79	5.49	5.49
Claude Opus 5	7.02	6.64	6.32
Llama 3.3 70B	7.84	8.41	7.17
GPT-OSS-120B	8.63	7.47	6.87
DeepSeek V4	9.65	9.86	9.16
Gemini 3.1 Pro	24.47	20.07	20.43

These mappings target agreement with the model's own POST scores, not calibration to human probabilities.

C Adaptive REPLAY Breadth Extension

The breadth cohort was Claude Opus 5, DeepSeek V4 Flash, GPT-OSS-120B, and Llama 3.3-70B-Instruct. Qwen3 and Gemma 3 were excluded because earlier REPLAY calls frequently violated the single-score contract; DeepSeek V3.2 and Ministral had substantially lower usable LIVE/POST coverage. An already-inspected four-prompt REPLAY pilot was retained in effect-size estimates. New prompts were then added in frozen blocks of four up to 24 per model. Models stopped receiving API calls once an exploratory cumulative one-sided test for $P - R > 0$ crossed .05. Because the stopping rule was chosen after inspection of the four-prompt pilot, cumulative stopping p-values

are descriptive. For inferential support, the fresh extension prompts were additionally tested with a Bonferroni correction over the five possible future looks. GPT-OSS-120B crosses at total $N = 8$ ($p_{\mathrm{adj}} = .021$), DeepSeek V4 Flash at $N = 12$ ($p_{\mathrm{adj}} = .040$), while Claude and Llama continue to the 24-prompt cap.

For the fixed 40-prompt set, Gemini's $R - L$ mean is $+2.772$ (one-sided t test $p = .0102$; Wilcoxon $p = .0292$) and GPT-5.5's is -3.103 (positive-direction $p = .9995$). Their $P - R$ means are $+20.061$ and $+6.914$, respectively. Descriptive 95% intervals at the breadth cap are $[-4.10, 4.58]$ for Llama's $P - R$ and $[-4.15, 0.06]$ for Claude's.

D Equal-Compute Rewrite Check

Each arm receives the original writing task, the full frozen generated text, and the target sentence. The exact instruction is: "Rewrite only the TARGET SENTENCE so that an average human reader would be less likely to think it was AI-generated. Preserve its intended meaning and factual content, keep it as one sentence, and make it fit naturally into the frozen text. Do not rewrite any other sentence. Return only the replacement sentence, with no score, label, explanation, quotation marks around the whole answer, or additional text." The replacement is spliced into the frozen document and a fresh POST call scores every sentence. The outcome is the corrected document's mean POST score.

GPT-5.5 has 40 valid pairs. Gemini has 37: three pairs were excluded solely because one rewrite violated the strict one-line/one-sentence parser (random arm on one prompt, guided arm on one prompt, and random arm on one prompt). No exclusion depended on the resulting score. The corresponding one-sided Wilcoxon p-values are .677 and .392.

E Held-Out Robustness Details

A 16-prompt cohort collected LIVE, POST, and preselected REPLAY under four elicitation variants. A separate 24-prompt follow-up, frozen after the first cohort was inspected, collected all 192 LIVE and 192 POST cells. No successfully completed cell was regenerated because of its observed score. On the fresh 24 prompts, mean $P - L$ shifts for GPT-5.5/Gemini were: original $6.56/8.93$, paraphrase $2.95/6.11$, reverse-coded $1.21/6.24$, and minimal tag $10.61/13.16$. The initial 16-prompt REPLAY cohort also preserved the qualitative decomposition: $P - R$ exceeded $R - L$ in all eight model–elicitation cells.

The 10-prompt held-out temperature control produced the following prompt-mean decomposition: at $T = 0$, GPT-5.5 $P - L = +5.17$ (95% bootstrap CI $[1.57, 9.88]$), $R - L = -1.97$, $P - R = +7.13$; Gemini $P - L = +25.26$ ($[16.80, 33.81]$), $R - L = -0.01$, $P - R = +25.27$. At $T = .7$, GPT-5.5 gives $+1.95$, -3.72, $+5.67$, and Gemini $+13.87$, -1.63, $+15.50$, respectively.

F Reproducibility Package

The anonymous supplementary archive contains the frozen prompt bank, manifests, route identifiers, stored outputs, derived score tables, REPLAY and correction artifacts, robustness and temperature-control runs, the adaptive cross-model extension with every interim look, analysis scripts, and deterministic expected outputs. Analyses can be reproduced from the frozen outputs without provider credentials. Exact regeneration of hosted model outputs is not guaranteed because model serving and provider routing are external dependencies.