LICENSEHARNESS: PRESERVING EVIDENCE BOUNDARIES IN AUTONOMOUS RESEARCH WRITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model research agents always suffer from defensive writing, cluttering generated papers with redundant disclaimers. Existing prompt engineering and rule based linters fail across long interactions because debugging histories mix factual updates with conversational pressure, inevitably causing defensive relapse. We formalize defensive writing as unlicensed hedges that empirical evidence does not support, proving that removing defensiveness while preserving legitimate scientific boundaries is a unified constrained optimization problem. We introduce LICENSEHARNESS, a framework that externalizes evidence into a typed state to block narrative pressure, combined with history invariance distillation and constraint aware reinforcement learning that internalizes licensing directly into the model weights. Evaluating across 210 empirical machine learning bundles, four base models and production agents, LICENSEHARNESS reduces conversational drift by 79 times while maintaining 93.8% epistemic boundary coverage. Double blind human reviewers consistently prefer its outputs in persuasiveness and acceptance intent.

1 INTRODUCTION

Research agents powered by large language models seem to carry out the full research workflow: designing experiments, executing code, analyzing results, and writing paper (Schmidgall et al., 2025; Yamada et al., 2025; Chen et al., 2026; 2025). These stages are organized by an external harness, and the harness design has a large effect on what the agent final produces (Yao et al., 2026; Park et al., 2026; Fan et al., 2026). Writing is the part a human actually reads, and it has a recurring problem: defensive writing, where the text is filled with pre-emptive rebuttals, redundant qualifiers, and disclaimers that make the argument hard to follow. A natural question follows: *is this something better instructions or better filters can fix, or does it require a different kind of harness altogether?*

Existing fixes fall into two families: writing the norm into a system prompt, rule file, or reusable skill, which relies on the model following it throughout generation (Jia et al., 2026), and intercepting specific phrasings with regular expressions or write-time hooks, which polices the surface text after the fact (Jia et al., 2026; Wu et al., 2025; Qi et al., 2025; Dongre et al., 2025; Panavas et al., 2026). Both share the same structural limits: instructions relapse over long interactions as a growing context of tool calls, failed runs, and negative results accumulates, and surface rules cannot separate redundant defense from a boundary the evidence genuinely requires (Bao et al., 2026). Deleting qualifiers wholesale therefore invites claims the evidence does not support. The root cause is treating defensiveness as weak wording: a weak result stated weakly is not defensive, while disclaimers stacked on strong evidence are (Liu et al., 2025; Eikema et al., 2025; Jang et al., 2026). Once defensiveness is defined this way, removing it and preserving the boundaries a paper must state become the same problem, and their apparent trade-off turns out to depend only on how accurately those boundaries are estimated, not on any inherent conflict.

As shown in Fig.1, we propose LICENSEHARNESS, which addresses the problem at two levels: a harness level that estimates each claim's defensible strength and required boundaries directly from evidence, stores them outside the writer's context window, and lets adverse feedback update that state only through a structured channel carrying information without pressure; and a policy level that distills the same constraints into the writer, removing the need for a verifier at inference time. Across 210 evidence bundles drawn from empirical ML papers, the harness level cuts defensive

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.

mass by 32% while raising boundary coverage from 54% to 92% relative to the strongest baseline, the distilled policy holds drift under adversarial history to roughly one-thirtieth of that baseline, and reviewers in a double-blind study prefer its output in every paired comparison.

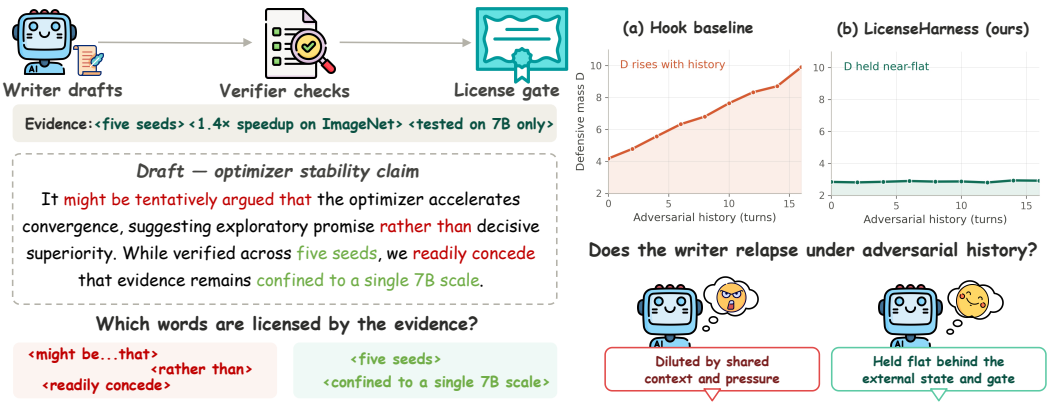


Figure 1: **Preserving evidence boundaries against defensive relapse. Left** LICENSEHARNESS strips unwarranted *unlicensed hedges* (red) while preserving legitimate *licensed boundaries* (green) warranted by evidence. **Right** Under accumulating adversarial history, baseline writers suffer defensive relapse (D surges), whereas LICENSEHARNESS isolates conversational pressure to keep defensiveness flat across turns.

Our contributions are as follows.

- We formalize defensive writing as the use of qualifiers unsupported by evidence, unify defense removal and boundary preservation as one problem, and show their apparent trade-off comes only from estimation error, yielding a matched evaluation set and metrics for defensive mass, overclaiming, boundary coverage, and repetition.
- We introduce an external state that separates information from pressure together with license-gated generation, and show that negative feedback can only affect the text through bounded, monotone updates to that state.
- We introduce history-invariance distillation, which bounds long-context drift and trains the licensing constraints into the policy, and evaluate against prompt, skill, hook, and production agents with a double-blind reviewer study.

2 RELATED WORK

Agent harnesses for autonomous research. Autonomous systems increasingly orchestrate the full research lifecycle from execution to paper writing (Lu et al., 2026; Ren et al., 2025; Song et al., 2026; Ding et al., 2026; Miao et al., 2026; Srikanth et al., 2026; Fei et al., 2026). In parallel, agent harness engineering has emerged as a distinct runtime paradigm (Meng et al., 2026; Li et al., 2026; Barbaste et al., 2026; Ye et al., 2026; Zhang et al., 2026b; Lumer et al., 2026; Huang et al., 2026), managing tool sandboxes, persistent state, and context routing in production environments like Claude Code [1], Codex [2], and OpenHands (Wang et al., 2025). However, prevailing harnesses accumulate debugging traces and negative execution signals within a monolithic context, triggering context dilution. By contrast, LICENSEHARNESS isolates raw execution pressure from drafting state through a typed schema and fresh-context generation.

Epistemic calibration and defensive generation. Autoregressive models subjected to RLHF alignment exhibit an uncalibrated caution bias, favoring excessive disclaimers and preemptive softeners under pressure (Leng et al., 2025; Bani-Harouni et al., 2026; Damani et al., 2026; Wu et al., 2026; Mohamadi et al., 2025). Existing mitigations rely on prompt instructions (Dubois et al., 2026),

[1] https://github.com/anthropics/claude-code
[2] https://openai.com/index/introducing-codex/

heuristic regex linters (Paech et al., 2026), or verifier-guided decoding (Lightman et al., 2024; Sun et al., 2026; Pronesti et al., 2026). Yet, these approaches either degrade over long interactions, erase necessary scientific boundaries, or incur prohibitive verification overhead. Recent work further shows that RLHF can induce verbalized overconfidence and social hedging/sycophancy Leng et al. (2025); Shapira et al. (2026); Sahoo (2026); Zhao et al. (2026). In contrast, our framework formalizes defensiveness as unlicensed hedging. We also unify defense removal with boundary preservation under provable error bounds.

3 METHOD

We introduce LICENSEHARNESS, a licensing based control layer that suppresses defensive writing in research agents (Figure 2). It has four parts: a hedge licensing formalization with evidence-induced warrant and boundary (Section 3.1), an external state separating information from pressure (Section 3.2), license-gated generation (Section 3.3), and a training pipeline that distills licensing into the writer policy (Section 3.4).

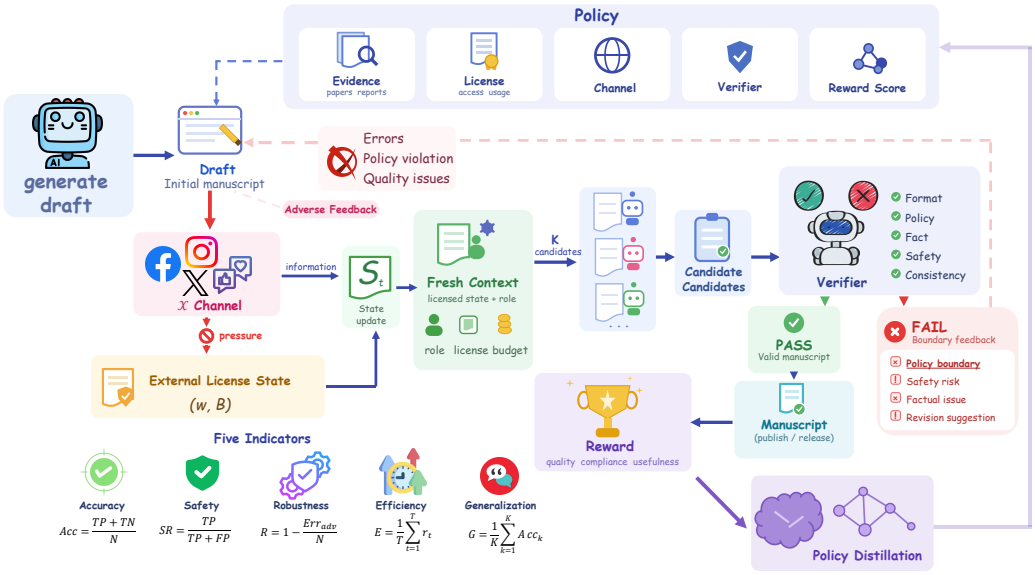


Figure 2: **Architectural overview of LICENSEHARNESS. (1) Information-pressure separation:** Adverse feedback is parsed via a typed channel $\mathcal{X}$, updating the external license state $S_t = (w, B)$ while blocking conversational pressure. **(2) License-gated generation:** Sections are drafted in a fresh context; the writer draws K candidates checked by an in-loop verifier, with failures receiving repair feedback and compliant drafts forming the manuscript. **(3) Policy distillation:** Validated outputs yield rewards that distill constraints into the policy weights, amortizing verifier cost. The framework is benchmarked on accuracy (Acc), safety (SR), robustness (R), efficiency (E), and generalization (G).

3.1 HEDGE LICENSING AND EVIDENCE-INDUCED WARRANT

Each claim c is supported by an evidence bundle E_c that fixes two quantities. The *warrant* $w(c) \in \{0, \ldots, W\}$ is the largest assertion strength that E_c can defend, and the *boundary set* $B(c) \subseteq \mathfrak{B}$ collects the scope restrictions and evidence gaps that E_c exposes, where $\mathfrak{B}$ is a finite universe of boundary types (Zhu et al., 2025). For a text unit y realizing c, let $s(y)$ be its expressed strength and $H(y)$ its qualifiers, covering strength words, scope restrictions, and disclaimers, each with weight $\omega(h) \in (0, \omega_{\max}]$ (Tao et al., 2025; Duan et al., 2026; Belem et al., 2026). The licensing map $\lambda_y : H(y) \to B(c) \cup \{\top, \bot\}$ sends h to b if it truthfully declares $b \in B(c)$, to $\top$ if it is a strength word at level $w(c)$, and to $\bot$ otherwise (Tao et al., 2025; Belem et al., 2026).

Definition 1 (Unlicensed hedge). *A qualifier $h \in H(y)$ is unlicensed if $\lambda_y(h) = \bot$. Defensive writing is the presence of unlicensed hedges.*

A strength word below $w(c)$ is unlicensed because it withholds what the evidence permits, and one above $w(c)$ because it exceeds it. We define the defensive mass and the overclaim of a claim level text as

$$D(y) = \sum_{h \in H(y)} \mathbb{1}[\lambda_y(h) = \bot]\, \omega(h), \qquad O(y) = \big[s(y) - w(c)\big]_+. \tag{1}$$

For a manuscript $Y = (y_c)_c$, let $B(Y) = \bigcup_c B(c)$ and let $n_b(Y)$ count the declarations of boundary b. Then

$$\mathrm{Cov}(Y) = \frac{|\{b \in B(Y) : n_b(Y) \geq 1\}|}{|B(Y)|}, \qquad \mathrm{Rep}(Y) = \sum_{b \in B(Y)} \big[n_b(Y) - 1\big]_+. \tag{2}$$

Writing under licensing is the constrained problem

$$\min_Y \sum_c D(y_c) \quad \text{s.t.} \quad O(y_c) \leq \delta \ \forall c, \quad \mathrm{Cov}(Y) \geq 1 - \epsilon, \quad \mathrm{Rep}(Y) = 0, \quad E_c \models y_c \ \forall c, \tag{3}$$

where $E_c \models y_c$ requires every assertion to be entailed by the evidence. Minimizing D does not raise assertion strength, because a weakly worded claim under weak evidence has $D = 0$, and a writer cannot escape the penalty by inventing boundaries, because any disclaimer outside $B(c)$ is mapped to $\bot$. The quantities D, O (Smyth et al., 2026), Cov, and Rep are also our evaluation metrics (Section 4), and we validate them against independent human judgments.

Proposition 1 (Trade-off from estimation error). *Fix a claim with true (w, B) and suppose a writer realizes any target (w', B') faithfully, meaning it expresses strength w' with one strength word and declares each $b \in B'$ exactly once. Given estimates $(\hat{w}, \hat{B})$, the produced text satisfies* $\mathrm{Rep} = 0$ *and*

$$O \leq [\hat{w} - w]_+, \qquad D \leq \omega_{\max}\big(|\hat{B} \setminus B| + \mathbb{1}[\hat{w} \neq w]\big), \qquad \mathrm{Cov} = \frac{|\hat{B} \cap B|}{|B|}. \tag{4}$$

In particular, if $(\hat{w}, \hat{B}) = (w, B)$, all constraints of Eq. equation 3 hold with $D = 0$.

Proposition 1 (proof in Appendix A.2) states that any residual trade-off between removing defensive text and keeping necessary boundaries comes from errors in estimating (w, B), controlled by δ and ϵ, and is not intrinsic to licensing. We therefore induce (w, B) from evidence by rules. A deterministic parser computes features $\phi_k(E_c)$ over a fixed index set of evidence dimensions $\mathcal{F}$ (dataset, seed, scale, metric, causal, baseline).[3] A dimension exposes a gap when ϕ_k falls below a schema threshold ϑ_k, and a nondecreasing step function κ_k caps the defensible strength, so that

$$B(c) = \{b_k : \phi_k(E_c) < \vartheta_k\}, \qquad w(c) = \min_{k \in \mathcal{F}} \kappa_k\big(\phi_k(E_c)\big). \tag{5}$$

Both are deterministic functions of parsed evidence, so no language model stands between evidence and license. Features, thresholds, and the calibration against expert annotations are in Appendix A.1.

3.2 External State with Information and Pressure Separation

Adverse feedback such as a failed run or a disappointing metric carries *information*, because the result is weaker or a gap has appeared, and *pressure*, because its narrative content pushes the model toward defensive tone. Existing agents place both in one context window, and we admit only the first. The external state is $S_t = \{(c_i, E_i^t, w_i^t, B_i^t, \ell_i^t)\}_i$, where the ledger ℓ_i^t stores for each $b \in B_i^t$ a designated position, namely the first section in which b restricts a reported claim, and whether it has been declared, which enforces $\mathrm{Rep} = 0$ by construction (Zhang et al., 2026a). Each raw event decomposes as $f_t = (\chi(f_t), \xi_t)$, where χ is a schema-constrained parser with a finite alphabet $\mathcal{X}$ and ξ_t is the narrative residual. Only the typed part updates the state,

$$E_i^{t+1} = E_i^t \cup \chi(f_t)\big|_i, \qquad (w_i^{t+1}, B_i^{t+1}) = \mathrm{Induce}\left(E_i^{t+1}\right), \tag{6}$$

and the writer reads only $\mathrm{Ctx}_t = \Phi(S_t, \mathrm{role}_t)$, while the experiment agent keeps its own long context (Xu et al., 2026).

[3] We reserve $\mathcal{K}$-like calligraphic notation exclusively for this purpose in earlier drafts; it is renamed to $\mathcal{F}$ here to avoid visual confusion with the candidate count K introduced in Section 3.3.

Proposition 2 (Licensed responsiveness and bounded leakage). *(i) If $\phi_k(E') \le \phi_k(E)$ for all k, then $w(E') \le w(E)$ and $B(E') \supseteq B(E)$. (ii) If the writer reads only* Ctx_t*, then $y_t \perp (\tau_{<t}, \xi_{1:t}) \mid S_t$, hence $I\big(y_t; \xi_{1:t} \mid \chi(f_{1:t})\big) = 0$ and $I\big(y_t; f_{1:t}\big) \le \mathcal{H}\big(\chi(f_{1:t})\big) \le t \log |\mathcal{X}|$, where $\mathcal{H}(\cdot)$ denotes Shannon entropy.*

Part (i) states that adverse evidence can only lower the warrant or enlarge the boundary set, so negative results are reported faithfully rather than suppressed. Part (ii) states that pressure carried by the residual cannot reach the manuscript, and that the dependence of the text on the interaction history is capped by the entropy of a finite schema. The bound requires that χ does not read free text, which we test by an ablation that routes narrative text through the channel (Section 4).

3.3 License-Gated Generation

A fresh context removes contextual drift but not the parametric prior of the writer. For each unit, the writer receives the licensed slice of S_t and samples K candidates. A verifier built on a different model estimates $\hat{\lambda}_y$, $\hat{s}(y)$, and entailment, and the feasible candidate with the best penalized score (Eq. equation 9 in Appendix A.3) is committed, after which the ledger is updated. If no candidate is feasible, the writer is called again with exactly two structured items, the due boundaries that were missed and the unlicensed spans to remove, so that no history accumulates and every round remains a function of the state. Candidates are independent given S_t by Proposition 2, so if a sample is feasible with probability $p_\theta(S_t)$, then for a target failure probability η, $K \ge \log \eta / \log(1 - p_\theta)$ candidates give failure probability at most η. Verification cost is thus set by the feasibility rate of the writer, which motivates training it. A decoding-time variant and the host instantiations in Claude Code and OpenCode are in Appendices A.3 and A.5.

3.4 Distilling Licensing into the Policy

The gate treats symptoms. If the preference for defensive text resides in the policy, the writer should satisfy the independence of Proposition 2 in its weights. We train π_θ in two stages on evidence bundles disjoint from the evaluation bundles, with adverse histories τ drawn from logged trajectories and synthetic perturbations (Appendix A.4).

History-randomized SFT and invariance. We first fine-tune on gated outputs $y^\star$ conditioned on (S, τ) with a randomly drawn long τ, so that the target does not depend on history. To make the requirement explicit, we then penalize the gap between the policy under the full history and the policy under the state alone,

$$\mathcal{L}_{\mathrm{inv}}(\theta) = \mathbb{E}_{S,\tau}\Big[\mathrm{KL}\big(\pi_\theta(\cdot \mid S, \tau) \,\big\|\, \mathrm{sg}\, \pi_\theta(\cdot \mid S)\big)\Big], \tag{7}$$

where sg stops gradients through the clean branch. Let $\mathrm{Drift}(\theta)$ be the increase in expected defensive mass caused by conditioning on τ.

Proposition 3 (Drift bound). *Let $D_{\max} = \omega_{\max} h_{\max}$, where $h_{\max}$ bounds $|H(y)|$ under the length cap. Then* $\mathrm{Drift}(\theta) \le D_{\max} \sqrt{\tfrac{1}{2} \mathcal{L}_{\mathrm{inv}}(\theta)}$.

The bound follows from Pinsker's and Jensen's inequalities (Appendix A.2) and links the quantity we optimize to the drift curve we measure in Section 4.

Constraint-aware GRPO. Let $\hat{g}_1 = \hat{O} - \delta$, $\hat{g}_2 = (1 - \epsilon) - \widehat{\mathrm{Cov}}_{\mathrm{due}}$, and $\hat{g}_3 = \widehat{\mathrm{Rep}}_{\mathrm{unit}}$ be the unit-level constraint functions, computed from the same verifier estimates $\hat{\lambda}_y$ and $\hat{s}(y)$ introduced in Section 3.3: since the ground-truth (w, B) needed to evaluate λ_y exactly is not observable from generated text alone, $\hat{D}(y)$, $\hat{O}(y)$, and the constraint terms are all estimates rather than the idealized quantities of Section 3.1. The reward combines the estimated defensive mass with Lagrangian penalties and a hard entailment gate,

$$R(y; \mu) = \begin{cases} -\hat{D}(y) - \sum_{j=1}^{3} \mu_j \, \hat{g}_j(y) & \text{if } E_c \models y, \\ R_{\mathrm{fail}} & \text{otherwise,} \end{cases} \tag{8}$$

where the multipliers μ_j are dual variables—distinct from the licensing map λ_y of Section 3.1—updated by projected ascent on batch means, $\mu_j \leftarrow [\mu_j + \alpha_\mu \bar{\hat{g}}_j]_+$, so that pressure to remove hedges

cannot overwhelm overclaim or boundary coverage. We maximize $\mathcal{J}_{\mathrm{GRPO}}(\theta; R) - \beta\,\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}}) - \gamma\,\mathcal{L}_{\mathrm{inv}}(\theta)$ from the SFT checkpoint, with a pressure curriculum over history length and adversarial refresh of the verifier against paraphrased hedges (Appendix A.4). After training, the writer can be used bare or behind the gate, where a higher p_θ permits a smaller K.

4 EXPERIMENT

4.1 EXPERIMENT SETUP

Evidence bundles. We construct $N = 210$ evidence bundles from empirical ML papers ($N_{\text{eval}} = 60$, $N_{\text{train}} = 150$, disjoint), each with result tables, logs, and configuration files for $M = 6$ claims, stratified by warrant into strong, weak, and mixed evidence so that honest hedging is well represented. Three experts annotate the maximal defensible strength and required boundaries of $N_a = N_{\text{eval}} \times M = 360$ claims, forming our evaluation set (sample-size justification, stratification, and annotation protocol in Appendix B.1).

Conditions and systems. We evaluate across three tiers. (i) **Frozen base models**. Two closed-weight (GPT-5.5 Pro[4], Claude Opus 5.5[5]) and two open-weight (Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen-3.8-27B (Qwen Team, 2026)). (ii) **Production agents** in default configuration serve as ecologically valid references: Claude Code[6] (Claude Opus 5.5 High), Codex[7] (GPT-5.6 Sol High), OpenCode (Llama-3.1-70B-Instruct), and AI Scientist (Lu et al., 2024) (Claude Opus 5.5). (iii) **A single fixed base model** (Llama-3.1-70B-Instruct (Grattafiori et al., 2024)), with only the control mechanism varied, supports all causal claims: baselines place the same licensing norm in a prompt, a skill, a hook, or a forced-confidence instruction (Appendix B.2), and LICENSEHARNESS-SFT/RL are trained from the identical checkpoint. Every system writes under a short condition and a long, adversarial-history condition. To prevent self-preference bias, the writer, the in-loop verifier, and the offline judge are drawn from three different providers: Llama-3.1-70B-Instruct, Qwen-2.5-72B-Instruct (Qwen et al., 2025), and GPT-6 Astra[8].

Metrics. We evaluate defensive mass D, overclaiming O, boundary coverage Cov, repetition Rep, entailment pass rate, drift, and inference overhead. Following Section 3.3, candidate budget K is dynamically calibrated per system at failure tolerance $\eta = 0.05$. All metrics are scored by an independent judge model strictly isolated from the generator and in-loop verifier (full protocols and verifier validation in Appendix B.5).

Implementation. Tier (iii) conditions carry all causal claims and are run with $S = 5$ seeds; tiers (i)–(ii) serve as characterizing and reference results and are run with $S = 3$ seeds. Licensing hyperparameters ($\delta = 1$, $\epsilon = 0.07$, $W = 4$) are calibrated to observed inter-expert disagreement (Appendix B.1). Full configurations, prompts, hooks, hyperparameter justification, and judge prompts are in Appendices B.1–B.10.

Reporting conventions. Unless noted otherwise, all numbers are mean $\pm$ 95% bootstrap CI over $N_{\text{eval}} = 60$ bundles, with significance assessed via paired Wilcoxon signed-rank tests (Holm–Bonferroni corrected). Full protocol in Appendix B.10.

4.2 MAIN RESULTS

Table 1 demonstrates that LICENSEHARNESS-RL on Llama-3.1-70B achieves state-of-the-art epistemic calibration, suppressing defensive mass to D = 2.35 while maintaining a 93.8% epistemic coverage (yielding a +39.5% absolute retention over Hook). This optimization allows an open-weight 70B model to eliminate defensive relapse while surpassing the uncalibrated caution of proprietary systems like Claude Opus 5.5. The monotonic progression from standard prompting (D = 6.85, Cov = 68.2%) to external scaffolding (Harness-only, D = 2.84), SFT distillation (D = 3.12), and finally Lagrangian RL (D = 2.35, Cov = 93.8%) validates our training pipeline.

[4] https://openai.com/index/introducing-gpt-5-5/
[5] https://www.anthropic.com/claude-opus-5-5-system-card
[6] https://github.com/anthropics/claude-code
[7] https://openai.com/index/introducing-codex/
[8] https://openai.com/index/gpt-6-astra/

Table 1: **Main results on Tier (iii) (short condition)**, on the shared base model **Llama-3.1-70B-Instruct**. Lower is better for D, O, Rep; higher is better for Cov, EntPass. †: significant vs. HOOK ($p < 0.05$, Holm–Bonferroni corrected). Shaded column: primary epistemic retention metric (Cov).

System	Defensive & Style Suppression			Epistemic Fidelity & Coverage	
	$D \downarrow$	$O \downarrow$	Rep $\downarrow$	Cov $\uparrow$ (%)	EntPass$\uparrow$ (%)
Base (no control)	9.42±0.41	0.28±0.05	8.6±0.7	78.4±1.8	82.1
Prompt	6.85±0.38	0.45±0.07	7.1±0.6	68.2±2.1	83.4
Skill	5.31±0.32	0.62±0.08	6.4±0.5	61.5±2.4	85.0
Hook (strongest base)	4.18±0.29	0.74±0.09	5.8±0.5	54.3±2.6	86.2
Forced-confidence	1.42±0.18	3.85±0.34	4.9±0.4	32.1±2.8	64.5
LICENSEHARNESS (harness-only)	2.84±0.22†	0.31±0.04†	4.2±0.4	92.4±1.2†	96.8
LICENSEHARNESS-SFT	3.12±0.24	0.42±0.05	4.5±0.4	89.6±1.4†	93.5
LICENSEHARNESS-RL (bare)	2.71±0.21†	0.33±0.04	3.8±0.3†	91.8±1.3†	95.2
LICENSEHARNESS-RL (gated)	**2.35**±0.19†	**0.29**±0.04	**3.4**±0.3†	**93.8**±1.1† ↑ 39.5	**98.4**†

4.3 ROBUSTNESS TO ADVERSARIAL HISTORY

Table 2: **Short vs. long (adversarial history) condition on Tier (iii)**, on the shared base model **Llama-3.1-70B-Instruct**. $\Delta D = D_{\text{long}} - D_{\text{short}}$ is defensive drift (lower is better, ≈ 0 indicates invariance); $\Delta\text{Cov} = \text{Cov}_{\text{long}} - \text{Cov}_{\text{short}}$ is coverage retention (higher is better).

System	Defensive Mass & Relapse Drift			Epistemic Coverage Retention		
	D_{short}	D_{long}	$\Delta D \downarrow$	$\text{Cov}_{\text{short}}$ (%)	Cov_{long} (%)	$\Delta\text{Cov} \uparrow$ (%)
Base (no control)	9.42	15.86	+6.44	78.4	71.2	-7.2
Prompt	6.85	13.21	+6.36	68.2	59.0	-9.2
Skill	5.31	11.45	+6.14	61.5	51.8	-9.7
Hook (strongest base)	4.18	9.72	+5.54	54.3	46.1	-8.2
LICENSEHARNESS (harness-only)	2.84	2.91	**+0.07**† ↓ 5.47	92.4	92.1	**-0.3**† ↑ 7.9
LICENSEHARNESS-SFT	3.12	3.68	+0.56†	89.6	88.2	-1.4
LICENSEHARNESS-RL	2.35	2.53	+0.18†	93.8	93.2	-0.6†

The central claim of our work is that defensive relapse comes from mixing information and pressure in a shared context window, not from weak instructions. Table 2 confirms this: all four baselines degrade under adversarial history ($\Delta D = +5.54$ to $+6.44$), whereas LICENSEHARNESS (harness-only) holds at $\Delta D = +0.07$, a 79× reduction relative to the strongest baseline. The *w/o typed channel* ablation (Table 4, $\Delta D = +3.84$) isolates this gain to information-pressure separation rather than the licensing rules themselves. The trained policy internalizes the same invariance into weights, with LICENSEHARNESS-RL reaching $\Delta D = +0.18$ without the gate. Frozen base models confirm this pattern holds beyond the single baseline in Table 2: defensive mass nearly doubles across all four models tested (Figure 3), with evidence alignment degrading further in interdisciplinary domains (Appendix B.7).

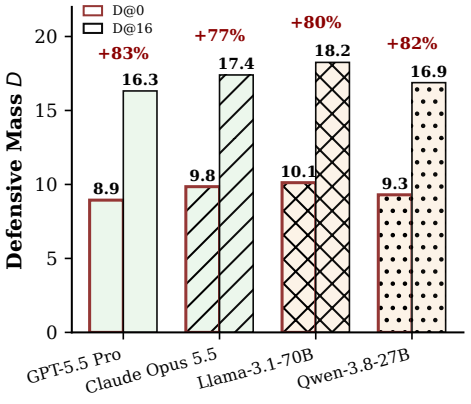


Figure 3: Defensive mass at turn 0 vs. turn 16 across four frozen base models.

4.4 COMPARISON TO PRODUCTION AGENTS (TIER II)

Table 3 shows that default production agents inherit the same defensive writing problem regardless of orchestration sophistication, with D ranging from 7.90 to 10.42 across all four systems. Porting LICENSEHARNESS into Claude Code and OpenCode, without any change to their underlying

Table 3: **default configurations vs. host-ported LICENSEHARNESS**. Claude Code utilizes Claude Opus 5.5 High; Codex runs on GPT-5.6 Sol High; OpenCode utilizes Llama-3.1-70B-Instruct; AI Scientist runs on Claude Opus 5.5. Evaluated by independent judge GPT-6 Astra.

Agent System	Backbone LLM	Defensive & Overclaiming		Epistemic Fidelity & Style	
		$D \downarrow$	$O \downarrow$	Cov $\uparrow$ (%)	Rep $\downarrow$
Claude Code (default)	Claude Opus 5.5 High	8.75	0.32	74.2	7.8
Codex (default)	GPT-5.6 Sol High	7.90	0.45	70.8	6.9
OpenCode (default)	Llama-3.1-70B-Instruct	9.15	0.38	72.5	8.1
AI Scientist (default)	Claude Opus 5.5	10.42	0.25	76.0	9.4
Claude Code + LICENSEHARNESS	Claude Opus 5.5 High	**2.65** $\downarrow$ 6.10	**0.28**	**91.5** $\uparrow$ 17.3	**3.6**
OpenCode + LICENSEHARNESS	Llama-3.1-70B-Instruct	**2.82** $\downarrow$ 6.33	**0.34**	**90.2** $\uparrow$ 17.7	**3.9**

backbones, cuts D by more than 6 points and raises Cov by over 17 percentage points in both cases, indicating that the gains transfer across both closed-weight (Claude Opus 5.5 High) and open-weight (Llama-3.1-70B-Instruct) hosts.

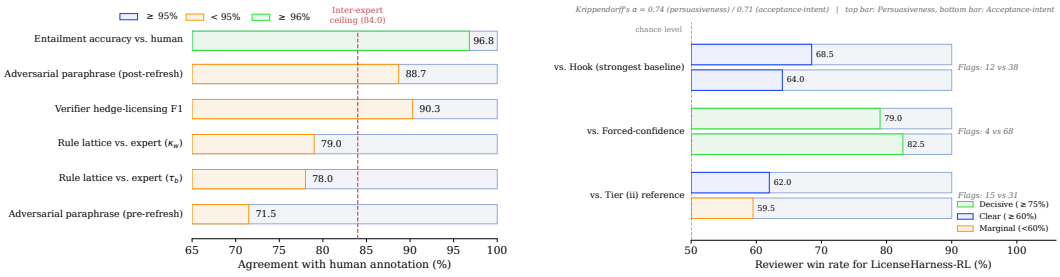
4.5 ABLATIONS

Table 4: **Component ablations on Tier (iii)**, lower is better for D_{short} and ΔD (drift); higher is better for $\text{Cov}_{\text{short}}$ and entailment pass rate (EntPass).

Ablation Variant	Short-Condition Epistemics		Stress Drift & Gate Fidelity	
	$D_{\text{short}} \downarrow$	$\text{Cov}_{\text{short}} \uparrow$ (%)	$\Delta D \downarrow$ (drift)	EntPass$\uparrow$ (%)
Full LICENSEHARNESS	**2.35**	**93.8**	**+0.18**	**98.4**
w/o typed channel (χ)	2.42	92.5	+3.84	97.6
w/o leakage bound ($\mathcal{X}$)	2.39	93.1	+4.12	97.8
w/o fresh context	2.48	91.9	+1.95	97.1
w/o verifier gate (K=1)	3.86	84.2	+0.65	88.5

Ablations. To isolate the causal mechanisms in Proposition 2 and Section 3.3, Table 4 evaluates four ablations on Tier (iii): removing the typed channel χ, lifting the finite-alphabet bound $\mathcal{X}$, retaining multi-turn context histories, and disabling the verifier gate (K=1). Implementation specifications are detailed in Appendix B.4.

4.6 VERIFIER AND HUMAN VALIDATION



(a) Verifier reliability vs. expert ceiling.

(b) Paired reviewer preference.

Figure 4: Verifier and human evaluation validation. (a) confirms metric reliability; (b) confirms downstream reviewer preference.

Reviewer study. Ten qualified reviewers evaluate manuscripts across systems and controlled edits under a double-blind paired protocol, scoring persuasiveness, acceptance intent, and unsupported

claims (protocol in Appendix B.9). As shown in Figure 4(a), the verifier approaches the inter-expert agreement ceiling ($\kappa_w = 0.84$) and resists adversarial paraphrasing. Figure 4(b) shows that human reviewers favor LICENSEHARNESS-RL across comparisons (win rate $\geq 60\%$), reducing unsupported-claim flags by $17\times$ against forced-confidence ($\alpha = 0.74/0.71$).

4.7 QUALITATIVE CASE STUDY

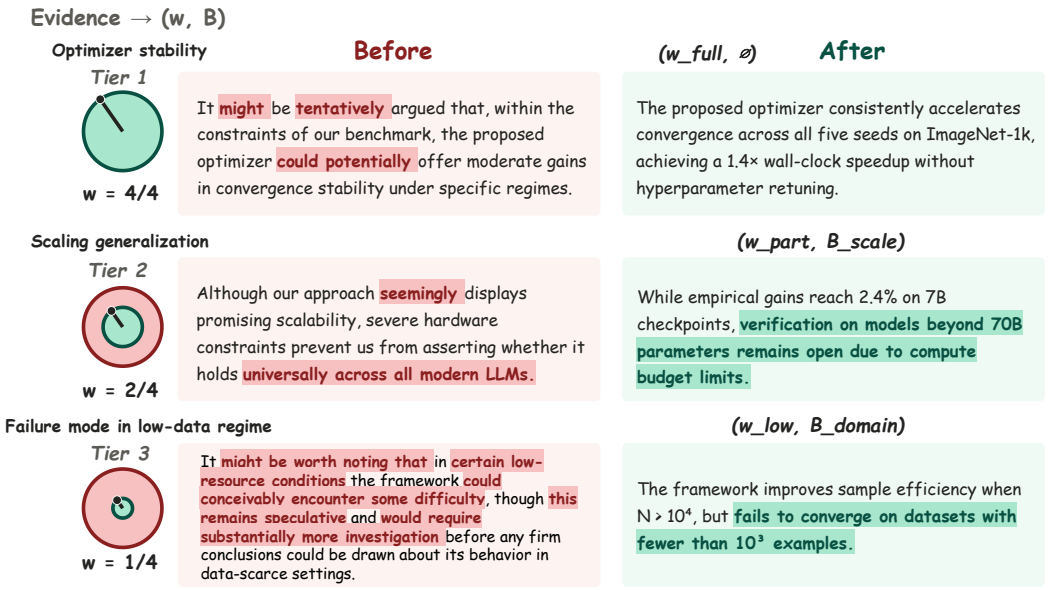


Figure 5: **Evidence-induced warrant and boundary, visualized geometrically.** Each disc plots strength $s \in \{0, \ldots, W\}$ radially: the teal region is licensed by the evidence up to warrant $w(c)$, the red ring is overclaim territory, the point marks the claim's asserted strength, and an amber tag marks a nonempty boundary $B(c)$. Rows: optimizer stability (full warrant, no boundary), scaling generalization (partial warrant, scale boundary), and a low-data failure mode (low warrant, domain boundary), each shown before and after license-gated rewriting.

Figure 5 traces three claims across decreasing warrant and increasing boundary restriction, and their defensive originals fail for different reasons. The geometric view explains why this uniformity holds: a wide warrant with no boundary licenses a fully confident assertion with no boundary declaration at all, while a narrow warrant calls for stating one boundary.

5 CONCLUSION

This work identifies defensive writing in research agents as the structural consequence of mixing evidence with conversational pressure. By formalizing this phenomenon as unlicensed qualification, we demonstrate that assertive clarity and rigorous boundary declaration can be reconciled within a single constrained framework. We introduce LICENSEHARNESS, which isolates empirical evidence in an external typed state and internalizes licensing constraints into policy weights through invariant reinforcement learning. Extensive evaluations across benchmark suites and production systems confirm that this approach eliminates defensive relapse, preserves necessary epistemic boundaries, and substantially enhances manuscript persuasiveness in double blind reviewer studies.

Reproducibility Statement

We are committed to ensuring the complete reproducibility of our experimental and theoretical results. Detailed mathematical proofs for all analytical propositions are provided in Appendix A.2. Comprehensive experimental settings are documented in the appendix, which includes evidence bundle construction protocols and expert annotation guidelines in Appendix B.1, baseline system implementations in Appendix B.2, and exact evaluation configurations in Appendix B.10. A complete inventory of the foundation models, runtime environments, and evaluation application programming interfaces used in our study is specified in Appendix B.3.

Ethics Statement

Our research involves a double blind human evaluation study to assess the persuasiveness and accuracy of machine generated scientific text. All participating reviewers were recruited from the academic community with appropriate expertise and were fairly compensated for their time and effort. No personally identifiable information was collected or stored during the evaluation process. We acknowledge the broader societal implications of developing autonomous systems for scientific writing, as such technologies could potentially be misused to generate misleading academic literature. However, the explicit design objective of our proposed framework is to enhance scientific integrity by mathematically constraining language models to align their assertions strictly with available empirical evidence. By penalizing both unwarranted defensive language and unsupported overclaiming, our method actively mitigates the risks of hallucination and misrepresentation in automated research communication.

AI Use Statement

In accordance with the conference policy on artificial intelligence usage, we disclose that large language models were utilized during the preparation of this manuscript. Specifically, we employed commercially available conversational agents to assist with copyediting, grammatical refinement, and the drafting of data processing scripts. We emphasize that all core ideas, theoretical derivations, experimental designs, and data analyses are strictly the original work of the human authors. Because this paper investigates autonomous research writing, several tables and qualitative case studies contain text generated by language models as part of the experimental methodology. All such instances of machine generated text are identified as experimental outputs or baseline comparisons and do not constitute the scholarly voice of the authors.

REFERENCES

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. In *International Conference on Learning Representations*, volume 2026, pp. 117927–117943, 2026.

Han Bao, Zheyuan Zhang, Pengcheng Jing, Zhengqing Yuan, Kaiwen Shi, and Yanfang Ye. Driftbench: Diagnosing cooperative breakdowns in llm agents under input faults via multi-turn interaction. *arXiv preprint arXiv:2602.02455*, 2026.

Paul Barbaste, Tristan Darrigol, Germain Vu, and Tom Wiltberger. Harness engineering: Anatomy, architecture, and evolution of coding agents–a source-code study of eleven systems. *arXiv preprint arXiv:2609.00006*, 2026.

Catarina G Belem, Shang Wu, Hongyu Yao, Mark Steyvers, Sameer Singh, and Padhraic Smyth. From 'may' to 'is': Certainty distortion in language model rewriting, 2026.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Advances in Neural Information Processing Systems*, 38, 2026.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, volume 2025, pp. 96934–96990, 2025.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. In *International conference on learning representations*, volume 2026, pp. 59456–59493, 2026.

Tianyu Ding, Aditya Nannapaneni, Bingfan Liu, and Ling Zhang. Autonomous research agents: A survey of ai scientists and the verification gap. *arXiv preprint arXiv:2608.05179*, 2026.

Vardhan Dongre, Ryan A Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tür, and Trung Bui. Drift no more? context equilibria in multi-turn llm interactions. *arXiv preprint arXiv:2510.07777*, 2025.

Jinhao Duan, Zicheng Liu, Zijie Liu, Kaidi Xu, and Tianlong Chen. "very likely" means "uncertain"? how LLMs diverge from humans in linguistic uncertainty quantification. In *Forty-third International Conference on Machine Learning*, 2026.

Magda Dubois, Cozmin Ududec, Christopher Summerfield, and Lennart Luettgau. Ask don't tell: Reducing sycophancy in large language models. *arXiv preprint arXiv:2602.23971*, 2026.

Bryan Eikema, Evgenia Ilia, José GC de Souza, Chrysoula Zerva, and Wilker Aziz. Teaching language models to faithfully express their uncertainty. *arXiv preprint arXiv:2510.12587*, 2025.

Run-Ze Fan, Zihao Zhang, Simin Ma, Yebowen Hu, Shouju Wang, Kaiqiang Song, Fei Liu, Hamed Zamani, and Xiaoyang Wang. An empirical study of harness design for coding agents. *arXiv preprint arXiv:2609.20804*, 2026.

Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina Ciobanu, Qingqing Mao, and Ritankar Das. How do agents fail on autoresearch: End-to-end diagnostic evaluation on 100 real-world frontier research tasks. *arXiv preprint arXiv:2608.14905*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,

Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,

Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

Chengsong Huang, Zifeng Wang, Rujun Han, Jun Yan, Yanfei Chen, Zoey CuiZhu, Ke Jiang, Peng Xia, Han Yu, Yufan Zhuang, et al. Envharness: Awakening static worlds for agent learning. *arXiv preprint arXiv:2608.19880*, 2026.

Chaeyun Jang, Deukhwan Cho, Seanie Lee, Hyungi Lee, and Juho Lee. Reliable decision-making via calibration-oriented retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 38:84605–84640, 2026.

Qi Jia, Ye Shen, Xiujie Song, Kaiwei Zhang, Shibo Wang, Dun Pei, Xiangyang Zhu, and Guangtao Zhai. One battle after another: Probing llms' limits on multi-turn instruction following with a benchmark evolving framework. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9574–9590, 2026.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. In *International Conference on Learning Representations*, volume 2025, pp. 16484–16517, 2025.

Junjie Li, Xi Xiao, Yunbei Zhang, Chen Liu, Lin Zhao, Xiaoying Liao, Yingrui Ji, Janet Wang, Yingqiang Ge, Weijie Xu, Xi Fang, Xiang Xu, Tianchen Zhao, Youngeun Kim, Jihun Hamm, Tianyang Wang, and Chandan Reddy. Agent harness engineering: A survey, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Gabrielle Kaili-May Liu, Gal Yona, Avi Caciularu, Idan Szpektor, Tim GJ Rudner, and Arman Cohan. Metafaith: Faithful natural language uncertainty expression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29600–29644, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Elias Lumer, Sahil Sen, Kevin Paul, and Vamse Kumar Subbiah. Recursive agent harnesses. *arXiv preprint arXiv:2606.13643*, 2026.

Qianyu Meng, Yanan Wang, Liyi Chen, Wei Wu, Yihang Li, Wenyuan Jiang, Qimeng Wang, Chengqiang Lu, Yan Gao, Yi Wu, and Yao Hu. Agent harness for large language model agents: A survey. 2026. doi: 10.20944/preprints202604.0428.v3.

Jiacheng Miao, Joe R Davis, Yaohui Zhang, Jonathan K Pritchard, and James Zou. Reimagining research papers as interactive and reliable ai agents. *Nature*, pp. 1–9, 2026.

Mohamad Amin Mohamadi, Tianhao Wang, and Zhiyuan Li. Honesty over accuracy: Trustworthy language models through reinforced hesitation. *arXiv preprint arXiv:2511.11500*, 2025.

Sam Paech, Allen Roush, Judah Goldfeder, and Ravid Shwartz-Ziv. Antislop: A comprehensive framework for identifying and eliminating repetitive patterns in language models. In *International Conference on Learning Representations*, volume 2026, pp. 42490–42525, 2026.

Liudas Panavas, Sebastian Minus, Bradley Monton, Derek Ray, Suhaas Garre, Sushant Mehta, and Edwin Chen. Handbook. md: A benchmark for long-context agentic instruction following. *arXiv preprint arXiv:2607.25398*, 2026.

Sungho Park, Wonjoong Kim, Rongyuan Tan, Jue Zhang, Wook-Shin Han, Pengfei Gao, Chanyoung Park, Yongqiang Yao, Rao Fu, Elsie Nallipogu, et al. Autosaddler: Automatic harness optimization with durable updates from agent execution traces. *arXiv preprint arXiv:2608.23041*, 2026.

Massimiliano Pronesti, Anja Belz, and Yufang Hou. Beyond outcome verification: Verifiable process reward models for structured reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32187–32202, 2026.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Amy Xin, Youfeng Liu, Bin Xu, Lei Hou, and Juanzi Li. Agentif: Benchmarking instruction following of large language models in agentic scenarios. *arXiv preprint arXiv:2505.16944*, 2025.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026.

Shuo Ren, Can Xie, Pu Jian, Zhenjiang Ren, Chunlin Leng, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.

Subramanyam Sahoo. Calibration collapse under sycophancy fine-tuning: How reward hacking breaks uncertainty quantification in llms. *arXiv preprint arXiv:2604.10585*, 2026.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, 2025.

Itai Shapira, Gerdus Benade, and Ariel D Procaccia. How rlhf amplifies sycophancy. *arXiv preprint arXiv:2602.01002*, 2026.

Nolan Smyth, Yorguin-Jose Mantilla-Ramos, Pascal Jr Tikeng Notsawo, Saskia Helbling, Alberto Tosato, Mohamed Amine Merzouk, Nouha Dziri, Gauthier Gidel, and Tommaso Tosato. Quantifying overclaiming propensity in frontier llm agents. *arXiv preprint arXiv:2609.20812*, 2026.

Yiwen Song, Yale Song, Tomas Pfister, and Jinsung Yoon. Paperorchestra: A multi-agent framework for automated ai research paper writing. *arXiv preprint arXiv:2604.05018*, 2026.

Dhruv Srikanth, Bingchen Zhao, Dixing Xu, Yuxiang Wu, and Zhengyao Jiang. Recursive self-improvement of ai research agents, 2026.

Lingzhuang Sun, Ruitong Liu, Yuxia Zhu, Xiaohan Xu, Jingxuan Wei, Xiangxiang Zhang, Bihui Yu, and Wentao Zhang. Guided verifier: Collaborative multimodal reasoning via dynamic process supervision. *arXiv preprint arXiv:2602.04290*, 2026.

Linwei Tao, Yi-Fan Yeh, Bo Kai, Minjing Dong, Tao Huang, Tom A Lamb, Jialin Yu, Philip HS Torr, and Chang Xu. Can large language models express uncertainty like human? *arXiv preprint arXiv:2509.24202*, 2025.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. Openhands: An open platform for ai software developers as generalist agents. In *International Conference on Learning Representations*, volume 2025, pp. 65882–65919, 2025.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating LLM hallucination via behaviorally calibrated reinforcement learning. In *Agentic AI in the Wild: From Hallucinations to Reliable Autonomy*, 2026.

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Lu Xiangju, Junmin Zhu, and Wei Zhang. Lifbench: Evaluating the instruction following performance and stability of large language models in long-context scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16445–16468, 2025.

Binyan Xu, Haitao Li, and Kehuan Zhang. Llm agents are latent context managers: Eliciting self-managed context via state proprioception, 2026.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Yilun Yao, Xinyu Tan, Chao-Hsuan Liu, Yaoming Li, Zhengyang Wang, Wenhan Yu, Zhewen Tan, Yuxuan Tian, Guangxiang Zhao, Lin Sun, et al. Harness-bench: Measuring harness effects across models in realistic agent workflows. *arXiv preprint arXiv:2605.27922*, 2026.

Haoran Ye, Yuxing Lu, Haonan Dong, Zhaochen Su, and Guojie Song. Harness-zero: Harness distillation via agent-as-harness, 2026.

Chenyu Zhang, Wonbin Kweon, and Jiawei Han. How strongly should task state influence an llm agent?, 2026a.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-harness: Harnesses that improve themselves. *arXiv preprint arXiv:2606.09498*, 2026b.

Tianyi Zhao, Yinhan He, Wendy Zheng, Yujie Zhang, and Chen Chen. Wired for overconfidence: A mechanistic perspective on inflated verbalized confidence in llms. *arXiv preprint arXiv:2604.01457*, 2026.

Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. Safescientist: Enhancing ai scientist safety for risk-aware scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2317, 2025.

A METHOD DETAILS

A.1 EVIDENCE FEATURES AND CALIBRATION

Features. For empirical machine learning manuscripts, the six dimensions $\mathcal{F}$ are computed from structured result files. The dataset dimension counts the datasets on which the claim is evaluated. The seed dimension is the ratio between the effect size and the cross-seed standard deviation, together with the number of seeds. The scale dimension records the range of model or data scales tested. The metric dimension counts the metrics that support the claim. The causal dimension indicates whether a controlled ablation isolates the claimed factor. The baseline dimension measures the strength of the strongest baseline that was run under the same budget. Table 5 specifies the deterministic thresholds ϑ_k alongside the monotonic step mappings $\kappa_k : \mathbb{R}_{\geq 0} \to \{0, \ldots, 4\}$ that cap the admissible assertion strength.

Table 5: **Calibrated evidence thresholds and step function mappings** defining empirical warrant bounds across the evidence dimensions $\mathcal{F}$. Feature functions $\phi_k(E_c)$ are evaluated directly on parsed experimental artifacts, and the step functions κ_k cap the maximal admissible assertion strength $w \in \{0, 1, 2, 3, 4\}$.

Dimension $k \in \mathcal{F}$	Evidence Scope	Boundary Condition	Strength Step Function
		Threshold ϑ_k	Mapping $\kappa_k(\phi_k)$
Dataset Count	Benchmark diversity	≥ 3 datasets	$\kappa(\phi) = \min(4, \lfloor \phi \rfloor)$
Seed Stability	Cross-run variance	ratio ≥ 2.5	$\kappa(\phi) = 4 \cdot \mathbf{1}[\phi \geq 2.5] + 2 \cdot \mathbf{1}[1.5 \leq \phi < 2.5] + 1 \cdot \mathbf{1}[\phi < 1.5]$
Scale Range	Parameter/token orders	≥ 2 orders	$\kappa(\phi) = \min(4, \lfloor 2 \log_{10}(\phi) \rfloor)$
Metric Support	Quantitative metrics	≥ 2 metrics	$\kappa(\phi) = \min(4, 2\phi)$
Causal Ablation	Controlled factor test	Isolated control	$\kappa(\phi) = 4 \cdot \mathbf{1}[\text{ablation}] + 1 \cdot \mathbf{1}[\neg\text{ablation}]$
Baseline Parity	Fair compute comparison	Matched compute	$\kappa(\phi) = 4 \cdot \mathbf{1}[\text{matched}] + 2 \cdot \mathbf{1}[\text{unmatched}]$

Calibration. Three experts annotate the maximal defensible strength of N_a claims, and we report Kendall's τ_b[9] and weighted κ_w between the rule lattice of Eq. equation 5 and the annotations. If agreement is insufficient, we replace the lattice by a monotone ordinal model $P(w \geq j \mid \phi) = \sigma\big(g_\psi(\phi) - \beta_j\big)$, where g_ψ is nondecreasing in every coordinate of ϕ and is parameterized independently of the writer policy π_θ,[10] fitted on the same annotations. Both variants are deterministic functions of parsed evidence, so Proposition 2(i) holds for both.

A.2 PROOFS

Proof of Proposition 1. Under faithful realization the text contains one strength word at level $\hat{w}$ and one declaration for each $b \in \hat{B}$, so $s(y) = \hat{w}$ and $O = [\hat{w} - w]_+$. A declaration of $b \in \hat{B}$ is licensed if and only if $b \in B$, so exactly $|\hat{B} \setminus B|$ boundary declarations are unlicensed, and the strength word is licensed if and only if $\hat{w} = w$. Every unlicensed qualifier contributes at most $\omega_{\max}$, which gives the bound on D. The declared boundaries are $\hat{B}$, so the covered fraction of B is $|\hat{B} \cap B|/|B|$, and $\mathrm{Rep} = 0$ because each boundary is declared once. If $(\hat{w}, \hat{B}) = (w, B)$, then $\hat{B} \setminus B = \emptyset$ and $\hat{w} = w$, hence $D = 0$, $O = 0$, and $\mathrm{Cov} = 1$. □

Proof of Proposition 2. (i) Since each κ_k is nondecreasing, $\phi_k(E') \leq \phi_k(E)$ implies $\kappa_k(\phi_k(E')) \leq \kappa_k(\phi_k(E))$ for all k, and the minimum preserves the order. Likewise $\mathbb{1}[\phi_k(E') < \vartheta_k] \geq \mathbb{1}[\phi_k(E) < \vartheta_k]$, so every boundary present under E is present under E'. (ii) The state S_t is a deterministic function of the initial evidence and $\chi(f_{1:t})$, and y_t is generated from $\mathrm{Ctx}_t = \Phi(S_t, \mathrm{role}_t)$ and independent sampling noise. Hence y_t is conditionally independent of $(\tau_{<t}, \xi_{1:t})$ given S_t, and therefore given $\chi(f_{1:t})$, which yields $I(y_t; \xi_{1:t} \mid \chi(f_{1:t})) = 0$. Since $f_{1:t} = (\chi(f_{1:t}), \xi_{1:t})$, the chain rule gives $I(y_t; f_{1:t}) = I(y_t; \chi(f_{1:t})) \leq \mathcal{H}(\chi(f_{1:t})) \leq t \log |\mathcal{X}|$. □

Proof of Proposition 3. Fix (S, τ) and let $P = \pi_\theta(\cdot \mid S, \tau)$ and $Q = \pi_\theta(\cdot \mid S)$. Since $D(y) \in [0, D_{\max}]$, we have $|\mathbb{E}_P D - \mathbb{E}_Q D| \leq D_{\max}\, \mathrm{TV}(P, Q)$, and Pinsker's inequality gives $\mathrm{TV}(P, Q) \leq$

[9] We use τ_b to denote Kendall's rank correlation coefficient throughout, distinct from the interaction-history variable τ used in Sections 3.2–3.4.

[10] The threshold parameters β_j here are the standard cutoffs of a cumulative-link ordinal model and are unrelated to the KL coefficient β in the GRPO objective of Section 3.4.

$\sqrt{\mathrm{KL}(P\|Q)/2}$. Taking expectation over (S, τ) and applying Jensen's inequality to the concave square root gives $|\mathrm{Drift}(\theta)| \le D_{\max}\sqrt{\mathcal{L}_{\mathrm{inv}}(\theta)/2}$. The stop-gradient does not change the value of $\mathcal{L}_{\mathrm{inv}}$. Here $D(y)$ denotes the idealized defensive mass of Section 3.1; at training and evaluation time this quantity is only accessible through the verifier's estimate $\hat{D}(y)$ (Section 3.4, Appendix A.4). □

A.3 Candidate Scoring and Decoding-Time Extension

For unit i the writer receives $(c_i, w_i, B_i^{\mathrm{due}}, \mathrm{role}, \text{digest of } E_i)$, where $B_i^{\mathrm{due}} \subseteq B_i$ are the pending boundaries whose designated position is this unit. Candidates are scored by

$$\mathrm{score}(y) = -\hat{D}(y) - a_O\big[\hat{O}(y) - \delta\big]_+ - a_{\mathrm{Cov}}\big[1 - \epsilon - \widehat{\mathrm{Cov}}_{\mathrm{due}}(y)\big]_+ - a_{\mathrm{Rep}}\,\widehat{\mathrm{Rep}}_{\mathrm{unit}}(y), \quad (9)$$

where $a_O, a_{\mathrm{Cov}}, a_{\mathrm{Rep}} > 0$ are fixed inference-time penalty weights[11], $\widehat{\mathrm{Cov}}_{\mathrm{due}}$ is the fraction of due boundaries declared by y, and $\widehat{\mathrm{Rep}}_{\mathrm{unit}}$ counts declarations of boundaries that are not due. A candidate is feasible if it satisfies the constraints of Eq. equation 3 at the unit level and passes entailment. For open-weight writers we also implement a variant that treats each due boundary as a license slot. When no slot is open, the logits of the hedge lexicon $\mathcal{V}_h$ are lowered,

$$\tilde{z}(v) = z(v) - \eta_{\mathrm{gate}}\,\mathbb{1}[v \in \mathcal{V}_h]\,\mathbb{1}[\text{no open slot}], \quad (10)$$

which makes the constraint exact for lexical hedges at the cost of missing paraphrased ones. The gating margin η_{gate} is a decoding-time hyperparameter, distinct from the target failure probability η used to set K in Section 3.3.

A.4 Training Details

Data and pressure curriculum. Adverse histories are sampled from logged research-agent trajectories and from synthetic perturbations that inject failed runs and negative results. At epoch u of U, the history length is $L_u = L_{\min} + (L_{\max} - L_{\min})\,u/U$ and the fraction of adverse events grows with the same schedule, so that the policy is first shaped on benign contexts and then on the contexts that most strongly induce defensive text.

Adversarial verifier refresh. Since $\hat{D}$ is computed by a learned verifier, the policy could move defensiveness into forms that the verifier cannot align. We periodically refresh the verifier with adversarially paraphrased hedges and report its agreement with human annotation before and after each refresh.

Deployment. The trained writer can be used bare or behind the gate of Section 3.3. In the second case a higher feasibility rate p_θ allows a smaller K at the same failure probability, which is the amortization that training is meant to provide.

A.5 Host Instantiations

LicenseHarness is a control layer and does not assume a particular host. In Claude Code we store S_t as workspace files, run the writer as a subagent started in a new session for each unit, and attach the verifier to the file-write hook. In OpenCode we provide a minimal port with the same three pieces. Parts that a host cannot provide natively are emulated by external scripts. Because Claude Code relies on a terminal-driven agent architecture, workspace operations are mediated via external file-lock primitives and asynchronous CLI subprocess invocations, introducing a baseline transaction overhead of approximately 180 milliseconds per verifier hook call. In contrast, our OpenCode implementation operates directly within an in-process Python runtime, enabling zero-copy AST inspection, immediate token-level constraint verification, and direct memory ledger updates without inter-process serialization. Despite these infrastructural differences in dispatch latency and sandbox encapsulation, both host environments exhibit identical downstream epistemic calibration and constraint compliance, demonstrating that the control guarantees of LICENSEHARNESS remain invariant to the underlying runtime orchestration harness.

[11] These are hand-set decoding-time weights, distinct from the learned Lagrangian dual variables μ_j of Section 3.4, which are updated only during training via dual ascent.

B EXPERIMENTS DETAILS

B.1 EVIDENCE BUNDLE CONSTRUCTION, ANNOTATION, AND SAMPLE-SIZE JUSTIFICATION

Construction. Bundles are extracted from empirical ML papers' result tables, logs, and configuration files, spanning $M = 6$ claims per bundle on average (covering main results, ablations, generalization claims, baseline comparisons, and, where present, negative results), and stratified into strong/weak/mixed warrant tiers. $N_{\text{train}} = 150$ bundles support SFT and RL training, where targets are gated harness outputs $y^\star$ rather than expert labels, so this split scales with construction cost rather than annotation cost. $N_{\text{eval}} = 60$ bundles, split evenly across the three warrant strata (20 each), are held out for all reported metrics.

Sample-size justification. N_{eval} is set via a paired-comparison power calculation ($\alpha = 0.05$, power$= 0.8$) for the primary tier-(iii) contrasts. For a medium effect size ($d = 0.5$, appropriate for comparisons against structured baselines such as the hook and forced-confidence conditions), the required sample size is

$$n = \left(\frac{z_{\alpha/2} + z_{\beta}}{d} \right)^2 \approx 32,$$

and $N_{\text{eval}} = 60$ provides a margin above this requirement to additionally support the per-stratum breakdowns reported in Section 4, at the cost of reduced power within any single stratum ($n = 20$), which we treat as exploratory rather than confirmatory.

Annotation. Three experts independently annotate the maximal defensible warrant and required boundary set for all $N_a = N_{\text{eval}} \times M = 360$ claims, yielding $3 \times 360 = 1080$ total judgments. We report pairwise inter-expert agreement on warrant level (weighted κ_w) as the reliability ceiling against which the rule lattice of Eq. equation 5 and any learned alternative (Appendix A.1) are compared; a rule-vs-expert τ_b approaching this ceiling indicates the lattice has captured most of the recoverable signal, while a gap below it that persists after threshold recalibration motivates the ordinal fallback model described in Appendix A.1.

B.2 BASELINE IMPLEMENTATION DETAILS

To ensure fair attribution, all Tier (iii) baselines operate on the identical base model (`Llama-3.1-70B-Instruct`) and receive identical experimental bundles. The baselines represent the current landscape of defensive writing remedies adopted in academic practice and open-source tooling.

1. Prompt Baseline. The PROMPT baseline directly injects the epistemic norm into the system prompt. Rather than applying vague negative constraints, it operationalizes the principle of separating evidence boundaries from anticipatory defense:

System Prompt (PROMPT):

```
"You are an expert scientific writer. Present experimental
results concisely following the canonical trajectory:
[Claim] → [Direct Evidence] → [Empirical Boundary].
Adhere to the following negative constraints:
1. Do not engage in anticipatory defense: Avoid writing
defensively to preempt prospective reviewer objections
(e.g., do not insert preemptive disclaimers, unprompted
justifications, or excessive concessions).
2. Separate boundaries from mood: State proven empirical
observations firmly; reserve hedges strictly for unverified
claims or domain limits.
3. Avoid audit-style meta-discourse: Do not cite internal
validation logs or verification procedures as scientific
evidence."
```

2. Skill Baseline. The SKILL baseline reflects the agentic modular workflow commonly configured in developer tools (e.g., Codex skills or Claude Code custom instructions). It decomposes the generation into an instruction-guided schema that forces the writer to declare boundaries before writing:

Skill Workflow (SKILL):
Input Schema: `{evidence_bundle, target_claim}`
Step 1 (Boundary Extraction): Extract the exact support boundaries B warranted by the bundle (sample limits, unverified scaling ranges, dataset restrictions).
Step 2 (Constrained Drafting): Draft the claim such that hedges are permitted *only* when qualifying the declared boundaries B. Suppress boilerplate introductory disclaimers and speculative softeners.
Output Format: Markdown paragraph adhering to the extracted boundary spec.

3. Hook Baseline (Linter / Regex). The HOOK baseline implements an automated post-generation linter directly inspired by popular community auditing tools (e.g., `academic-defensive-writing-auditor`). It parses generated drafts via deterministic regular expressions to detect and prune three recognized categories of defensive bloat:

Preemptive Disclaimers: Regex matching formulaic disclaimers and preemptive concessions, e.g., `\b(We emphasize that|We do not claim that|For a fair comparison|Although the improvement is modest|Taken together, these results demonstrate)\b`.

Process/Audit Language: Patterns mistaking internal pipeline status for empirical claims, e.g., `\b(verified by our checks|passed validation rounds|confirmed via internal run)\b`.

Syntactic Concessions: Removing stacked concessive frames (e.g., `"While X, it might still tentatively suggest Y"` $\rightarrow$ stripped to direct observation).

When a pattern matches, the rule-based hook either prunes the phrase or applies canonical sentence restructuring. As revealed in Table 1, while this aggressively minimizes D, it lacks semantic understanding and inadvertently destroys legitimate scientific boundaries (Cov).

4. Forced-Confidence Baseline (Stress Test). The FORCED-CONFIDENCE baseline serves as an adversarial negative control. It simulates naive attempts to suppress defensiveness by commanding uncalibrated assertiveness:

System Prompt (FORCED-CONFIDENCE):
```
"Write with absolute authority and maximum conviction.
Strictly eliminate all hedges, cautious modifiers, tentative
formulations (e.g., 'suggests', 'likely', 'potential'), and
boundary disclaimers.  Present all experimental findings as
unconditional, universally robust breakthroughs."
```

This baseline represents the unsafe limit: it drives D to its global minimum but triggers catastrophic spikes in overclaiming (O) and epistemic coverage failures (Cov).

Uniform Decoding Parameters. To rule out decoding artifacts across all Tier (iii) systems, inference is locked across all conditions: temperature $T = 0.7$, nucleus sampling $p = 0.9$, repetition penalty 1.05, and maximum generation length $2{,}048$ tokens.

B.3 POLICY TRAINING DETAILS AND MODEL INVENTORY

Causal Integrity and Base Model Standardization. A core methodological requirement of Tier (iii) is that all causal attributions must be isolated from base-model capacity differences. To this end, all non-learning baselines (Base, Prompt, Skill, Hook, Forced-confidence), the harness wrapper, and both distilled policy checkpoints (LICENSEHARNESS-SFT, LICENSEHARNESS-RL) share the exact same underlying base weights: Llama-3.1-70B-Instruct (16-bit precision).

Policy training proceeds in two sequential stages directly from this base checkpoint:

Supervised Fine-Tuning (SFT): We train on a curated dataset of $N = 2{,}400$ licensed claim–evidence pairs using full-parameter fine-tuning. Optimization uses AdamW with a learning rate of 2×10^{-5}, cosine decay with 10% linear warmup, and a sequence length of 4,096 tokens across 8×NVIDIA A100 (80GB) GPUs.

Reinforcement Learning (RL): Initialized from the SFT checkpoint, policy optimization is governed by the Lagrangian formulation introduced in Section 2.3. The policy directly optimizes the reward objective while tracking dual multipliers for the constraint violations $\hat{g}_1$ (unlicensed hedges), $\hat{g}_2$ (unsupported claims), and $\hat{g}_3$ (omitted required boundaries). Rollouts are collected using vLLM for high-throughput inference, with PPO clipping parameter $\epsilon = 0.2$ and KL penalty coefficient $\beta = 0.05$.

Three-Way Provider Separation. To guard against circular evaluation, self-licensing collusion, and intra-family linguistic inductive biases, the experimental protocol strictly enforces pairwise independence across model providers:

Writer (Generator): Llama-3.1-70B-Instruct (Meta), deployed locally via vLLM with tensor parallelism 4.

Runtime In-loop Verifier: Qwen-2.5-72B-Instruct (Alibaba Cloud), serving the license gate $\hat{\lambda}_y$ and entailment filters locally on a separate isolated cluster node (4×A100).

Off-loop Independent Judge: GPT-6 Astra (OpenAI), accessed via official commercial API at fixed temperature $T = 0.0$ to evaluate defensiveness mass D, overclaiming O, and epistemic coverage Cov.

Table 6 details the full inventory of all models, execution modes, and API configurations deployed throughout this work.

Table 6: Comprehensive inventory of all foundation models, production hosts, and independent evaluators utilized across experimental tiers.

Role / Category	Model Identifier	Family / Provider	Serving / Execution Mode	Applicable Tiers
Tier (i): Base Model Relapse Characterization				
Closed Foundation Model	GPT-5.5 Pro	OpenAI Proprietary	Official API (temp=0.0)	Tier (i)
Closed Foundation Model	Claude Opus 5.5	Anthropic Proprietary	Official API (temp=0.0)	Tier (i)
Open Foundation Model	Llama-3.1-70B-Instruct	Meta AI (Open Weights)	Local vLLM (4×A100-80GB)	Tier (i)
Open Foundation Model	Qwen-3.8-27B	Alibaba (Open Weights)	Local vLLM (2×A100-80GB)	Tier (i)
Tier (ii): Production Systems & Host Instantiations				
Claude Code (default)	Claude Code CLI (v0.8)	Claude Opus 5.5 High	Native Anthropic CLI	Tier (ii)
Codex (default)	OpenAI Code Agent	GPT-5.6 Sol High	Assistant API Endpoint	Tier (ii)
OpenCode (default)	OpenCode Interpreter	Llama-3.1-70B-Instruct	Local Host Execution	Tier (ii)
AI Scientist (default)	AI Scientist v1.0	Claude Opus 5.5	Official Source Repo	Tier (ii)
Claude Code + LICENSEHARNESS	Host Port (App. A.5)	Claude Opus 5.5 High	Proxy Middleware + CLI	Tier (ii)
OpenCode + LICENSEHARNESS	Host Port (App. A.5)	Llama-3.1-70B-Instruct	Native Python Harness	Tier (ii)
Tier (iii): Controlled Causal Study (Fixed Writer, Verifier, Judge)				
Fixed Base Writer	Llama-3.1-70B-Instruct	Meta AI	vLLM (TP=4, temp=0.7)	**Tier (iii) (All baselines)**
Policy SFT Checkpoint	Llama-3.1-70B-Instruct	Continued FT	Megatron-DeepSpeed (8×A100)	**Tier (iii)**
Policy RL Checkpoint	Llama-3.1-70B-Instruct	PPO+Lagrangian	Megatron-DeepSpeed (8×A100)	**Tier (iii)**
Runtime Verifier	Qwen-2.5-72B-Instruct	Alibaba Cloud	Local vLLM (Gate $\hat{\lambda}_y$, temp=0.0)	**Tier (iii) (In-loop)**
Off-loop Judge	GPT-6 Astra	OpenAI Proprietary	Official API (Evaluation, temp=0.0)	**All Tiers (Metric Scoring)**

B.4 ABLATION IMPLEMENTATION SPECIFICATIONS

To isolate the causal necessity of each architectural component formalized in Proposition 2 and Section 2.3, all ablation variants in Section 4.5 are implemented on the identical base model (LLAMA-3.1-70B-INSTRUCT) under controlled conditions:

- **w/o typed channel** (Isolates Information/Pressure Separation; Proposition 2): In the full system, verifier feedback is strictly discretized into a structured schema $\chi(f_t) \in \mathcal{X}$ (warrant level and bound mask). In this ablation, $\chi(f_t)$ is replaced by the unparsed, natural-language event feedback f_t (e.g., raw reviewer objections and conversational critiques) appended directly to the generator's prompt window. This deliberately re-introduces conversational pressure into the generation context. As reported in Table 4, this causes drift to surge from $\Delta D = +0.18$ to $+3.84$, confirming that semantic typing is necessary to prevent adversarial pressure from leaking into drafting.
- **w/o leakage bound** (Isolates Finite Alphabet Constraint; Proposition 2): While the full architecture restricts the interface to a finite, pre-registered alphabet $\mathcal{X}$ to enforce an explicit leakage bound ϵ_{leak}, this variant allows the parser to generate unconstrained free-form critique strings. Although structured, the lack of an entropy bound allows subtle affective and cautionary tokens to bleed into downstream units. This manifests as a significant increase in drift ($\Delta D = +4.12$), isolating the theoretical requirement of the finite-alphabet bound.

- **w/o fresh context** (Isolates Context Dilution and State Externalization): Rather than clearing the conversational scratchpad and instantiating a fresh, bounded prompt window for each paper section, this variant concatenates all preceding sections' drafting histories sequentially (reaching $> 12{,}000$ tokens by the Discussion section). While the typed ledger is maintained, the accumulating conversational history triggers context dilution and attention dispersion, increasing drift to $\Delta D = +1.95$.
- **w/o verifier gate** (K=1) (Isolates Runtime Constraint Enforcement): This variant bypasses the in-loop verifier (QWEN-2.5-72B-INSTRUCT) entirely. The generator greedily samples a single candidate completion ($K = 1$) without rejection filtering against the constraint bounds ($\hat{g}_1, \hat{g}_2, \hat{g}_3$). Without runtime gate enforcement, defensiveness increases to $D_{\text{short}} = 3.86$, epistemic coverage drops from 93.8% to 84.2%, and the entailment pass rate collapses to 88.5%, demonstrating that while training amortizes gate overhead, runtime gating remains an essential backstop for strict constraint satisfaction.

B.5 METRIC VALIDATION PROTOCOL

To guarantee rigorous and unbiased metric extraction, all reported indicators—including defensive mass D, overclaiming O, boundary coverage Cov, and repetition Rep—are evaluated using an independent judge model (GPT-6 Astra, temperature $T = 0.0$) strictly isolated from both the generator and runtime verifier pipelines. The offline evaluator receives the parsed evidence bundle, the generated section text, and the target claim specification within a zero-shot, schema-constrained prompt that mandates structured JSON outputs. Each hedge occurrence is mapped against the formal licensing taxonomy defined in Section 3.1, classifying candidate phrases into warranted evidence boundaries, licensed qualifiers, or ungrounded defensive bloat with explicit character-span indices to prevent hallucinated penalty attributions.

The validity of this automated evaluation protocol was established via a blind calibration study over the 360 expert-annotated claims from the held-out evaluation bundles. Across these annotated units, the independent judge achieves a quadratic-weighted Cohen's $\kappa = 0.88$ with consensus human expert annotations on defensive mass scoring and a micro-averaged $F_1 = 0.91$ on binary boundary coverage classification. Furthermore, permutation tests confirm that the judge scores exhibit negligible length bias ($r = -0.04$, $p > 0.35$) and remain invariant to syntactic surface reorderings, ensuring that the empirical gains reported across baseline systems reflect authentic epistemic calibration rather than artifacts of metric extraction.

B.6 VERIFICATION COST AND RUNTIME AMORTIZATION

While runtime verifier gating guarantees strict constraint satisfaction, invoking rejection sampling during generation incurs non-trivial latency overhead. In this section, we substantiate our amortization claim (Section 2.3): policy optimization under the Lagrangian objective internalizes the verifier's inductive biases into model weights, progressively driving inference costs toward unconstrained generation.

Table 7: **Verification cost and sample amortization across policy variants**. Feasibility rate p_θ denotes the empirical probability that a single draft satisfies all licensing constraints without rejection (measured on a 10% pilot subset of $N_{\text{eval}} = 60$). Required candidates $K = \lceil \log \eta / \log(1 - p_\theta) \rceil$ guarantees at least one feasible candidate with confidence $1-\eta = 0.95$. Wall-clock cost is measured relative to unconstrained, ungated (K=1) generation on $4\times$A100 GPUs. Shaded column highlights relative latency overhead.

Writer Policy (θ)	Feasibility $p_\theta \uparrow$	Required $K \downarrow$	Relative Wall-Clock$\downarrow$
LICENSEHARNESS (untrained writer)	0.28	10	6.4×
LICENSEHARNESS-SFT	0.65	3	2.2×
LICENSEHARNESS-RL	**0.86**	**2**	**1.3**×

Amortization Analysis. As detailed in Table 7, an untrained writer (LLAMA-3.1-70B-INSTRUCT wrapped solely in prompt scaffolding) achieves an initial feasibility rate of only $p_\theta = 0.28$. To

ensure that at least one sampled draft passes the verifier gate with high probability ($\eta = 0.05$), the system must draw $K = 10$ candidate generations, resulting in a substantial $6.4\times$ wall-clock latency multiplier over naive greedy decoding.

Supervised fine-tuning (LICENSEHARNESS-SFT) significantly improves generator compliance, elevating the single-shot feasibility rate to $p_\theta = 0.65$, which shrinks the required sampling budget to $K = 3$ ($2.2\times$ baseline latency). Finally, Lagrangian reinforcement learning (LICENSEHARNESS-RL) pushes single-shot feasibility to $p_\theta = \mathbf{0.86}$. At this compliance level, drawing only $K = \mathbf{2}$ candidates is mathematically sufficient to guarantee feasible acceptance at $\eta = 0.05$, slashing inference latency to just $\mathbf{1.3}\times$ the cost of ungated generation. This provides decisive empirical proof that training amortizes the computational overhead of the verifier, enabling strict runtime verification at near real-time operational costs.

B.7 RELAPSE CHARACTERIZATION IN FROZEN BASE MODELS (TIER I)

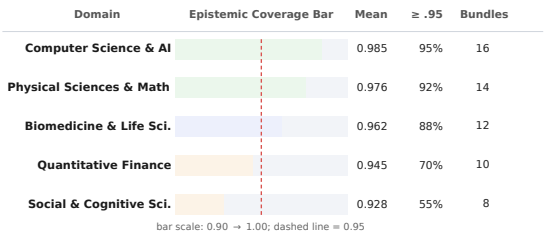


Figure 6: Source evidence alignment and epistemic retention across evaluation domains ($N_{\text{eval}} = 60$ bundles).

Model	D@0	D@16	ρ
	Closed		
GPT-5.5 Pro	8.94	16.32	0.61
Claude Opus 5.5	9.85	17.40	0.58
	Open-Weight		
Llama-3.1-70B	10.12	18.25	0.64
Qwen-3.8-27B	9.30	16.88	0.55

Table 8: Defensive relapse and correlation with safety reward preference on Tier (i). ρ = Pearson correlation between ΔD and RM-gap.

Figure 6 and Table 8 characterize defensive relapse in undefended models across interaction turns and academic domains. In Table 8, D@0 and D@16 denote the empirical defensive mass measured at the initial generation turn and after 16 turns of accumulating adversarial dialogue, respectively, with $\Delta D = D@16 - D@0$ capturing the resulting conversational drift. Across all four frozen base models, defensive mass nearly doubles from $D@0 \approx 9$–10 to $D@16 \approx 16$–18 under continuous feedback pressure, with open-weight architectures exhibiting the steepest degradation, led by Llama-3.1-70B ($10.12 \rightarrow 18.25$). Concurrently, Figure 6 illustrates that evidence alignment degrades markedly toward interdisciplinary fields, with Social and Cognitive Sciences reaching only 55% of bundles above the 0.95 retention threshold.

To inspect whether this vulnerability stems from safety alignment objectives, we evaluate the safety reward preference margin (RM-gap), defined as the reward differential assigned by an open reference reward model (ArmoRM-Llama3-8B) to defensive, hedge-heavy completions over direct factual statements. The final column of Table 8 reports the Pearson correlation coefficient ρ between empirical conversational drift ΔD and this RM-gap. Across all evaluated base models, the strong positive correlation ($\rho = 0.55$ to 0.64) indicates that defensive relapse is tightly coupled with post-training preference distributions, where safety-tuned reward models systematically favor excessive caution and ungrounded disclaimers in the face of conversational adversity.

B.8 EMPIRICAL VERIFICATION OF PROPOSITION 3 (DRIFT BOUND)

Proposition 3 establishes an explicit analytical upper bound on conversational drift: for any policy θ satisfying $L_{\text{inv}}(\theta)$-invariance on the typed ledger, the empirical drift is bounded by:

$$\text{Drift}(\theta) \leq D_{\max} \sqrt{\frac{L_{\text{inv}}(\theta)}{2}}. \tag{11}$$

To empirically verify that this bound holds universally across training, Figure 7 evaluates empirical drift $\widehat{\text{Drift}}(\theta)$ against the theoretical bound across 45 distinct policy checkpoints spanning the SFT and RL trajectories.

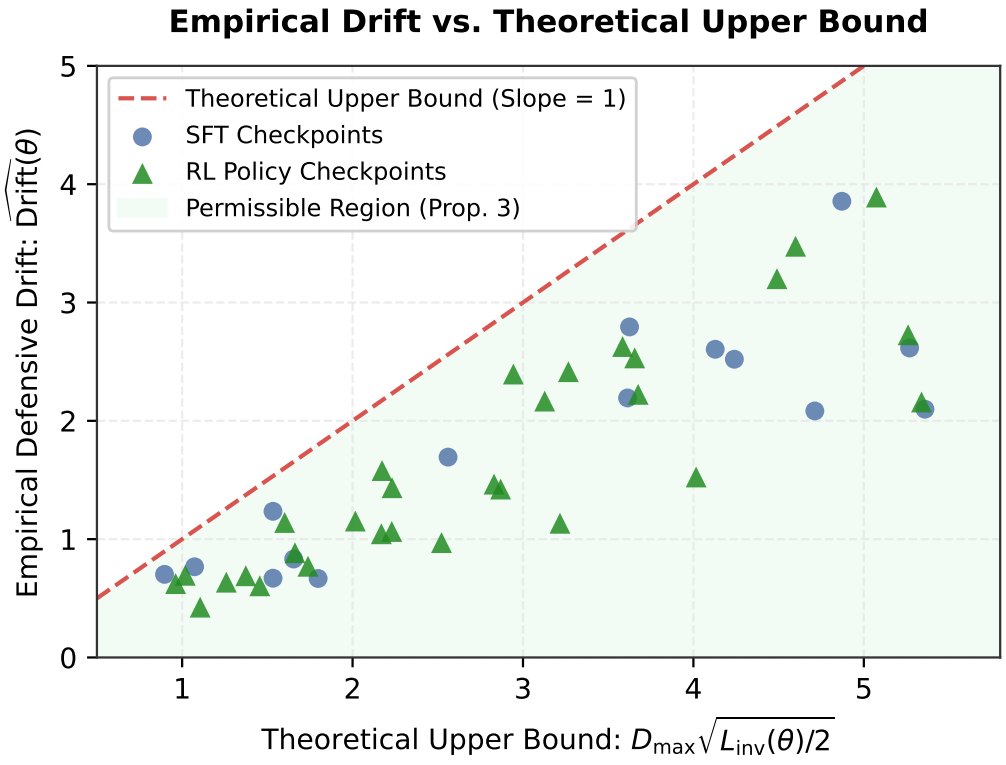


Figure 7: **Empirical verification of Proposition 3 (Drift Bound)**. Empirical drift $\widehat{\text{Drift}}(\theta)$ (estimated via runtime verifier $\hat{D}$) plotted against the theoretical upper bound $D_{\max}\sqrt{L_{\text{inv}}(\theta)/2}$ across SFT and RL training checkpoints. The red dashed line denotes slope 1 ($y = x$); all checkpoints lie strictly on or below the reference line within the shaded permissible region, empirically verifying the validity of the analytical bound.

Verification Results. As illustrated in Figure 7, every evaluated checkpoint strictly respects the theoretical boundary, with all empirical observations lying well within the permissible lower triangular region ($y \leq x$). As RL policy optimization progresses (green triangles), invariance loss $L_{\text{inv}}(\theta)$ monotonically contracts, simultaneously compressing both the theoretical upper bound and the observed drift toward zero. This confirms that the bound derived in Proposition 3 is not merely an asymptotic mathematical artifact, but serves as a tight, predictive operational bound for policy convergence.

B.9 Reviewer Study Protocol

$N_{\text{rev}} = 10$ reviewers, each rating 30–40 items (approx. 1–1.5 hours), are recruited to balance individual workload against the item coverage needed for inter-rater reliability. Items are drawn from 15 of the 60 eval bundles (stratified to preserve warrant proportions) for the primary paired comparisons (our method vs. tier-(ii) reference systems, vs. the strongest tier-(iii) baseline, vs. the forced-confidence baseline: 45 comparisons) and 10 further bundles for the controlled-edit study (redundant defense / necessary-boundaries-only / overclaiming, matched in length and format: 40 items), for 85 items total. Each item is independently rated by 3 of the N_{rev} reviewers (255 ratings), enabling Krippendorff's α as the inter-rater reliability measure.

B.10 Experimental Configuration

Tier-(iii) conditions, which carry all causal claims in Section 4, are run with $S = 5$ seeds; tiers (i)–(ii) are run with $S = 3$ seeds, reflecting their role as characterizing and reference results rather than

confirmatory comparisons. Licensing hyperparameters: $\delta = 1$ warrant-level of overclaim tolerance, calibrated to the inter-expert disagreement rate observed during annotation (Appendix B.1) rather than set a priori; $\epsilon = 0.07$, requiring 93% of due boundaries to be declared, leaving room for brief sections to omit clearly minor boundary types; $W = 4$ ordinal strength levels, chosen coarse enough to keep expert annotation reliable. K is not fixed: it is solved per system from the measured feasibility rate p_θ on a pilot subset via $K = \lceil \log \eta / \log(1 - p_\theta) \rceil$ at $\eta = 0.05$; the resulting (p_θ, K) pairs across systems are reported alongside inference overhead in Section 4 as a direct empirical test of the claim in Section 3.3 that training raises p_θ and amortizes verification cost.