PROMPT INJECTION AS AN INFORMATION CHANNEL: DETECTION LIMITS AND HIERARCHY CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt-injection evaluations often conflate three distinct questions: whether benign and attacked output laws are distinguishable, how much attacker-controlled information can flow through generation, and whether a finite audit can certify a hierarchy-violation bound. We give a unified information-theoretic framework that separates these questions and connects each to an auditable quantity. We define benign and attacked model-induced output laws under an explicit observation model and characterize their optimal one-shot testing risk. We then replace an independent-token analysis with an autoregressive likelihood-ratio theorem: under auditable drift and martingale-tail conditions, detection error decreases with accumulated conditional information. We prove that attention entropy alone cannot provide an architecture-independent injection-capacity bound, and give an influence-aware alternative. Finally, protected-event data processing yields a hierarchy-violation certificate whose finite-sample form uses exact binomial coverage under a declared IID audit-unit model rather than zero-rate clipping. Across 18,432 Kaggle trials, attack-conditioned exact leakage spans 0.137–0.590 across four open-model configurations, while the best pair-conditioned oracle test reaches held-out AUC 1.000 and output-surface baselines remain substantially weaker. A controlled hierarchy LoRA reduces direct-override leakage but does not yield a confidence-certified increase in required KL divergence, separating empirical robustness from certification.

1 INTRODUCTION

Prompt injection exploits a structural ambiguity in language-model systems: trusted policy, user instructions, and untrusted retrieved content are all represented as tokens consumed by one generator. Direct attacks override an application prompt (Perez & Ribeiro, 2022; Liu et al., 2024); indirect attacks place instructions in documents or tool outputs (Greshake et al., 2023; Zhan et al., 2024), and adaptive attacks bypass defenses tuned to fixed threat sets (Liu et al., 2025; Zhan et al., 2025). Instruction hierarchy, structured queries, and preference optimization improve robustness but do not establish an absolute boundary (Wallace et al., 2024; Chen et al., 2025a;b; Geng et al., 2025).

Existing formalizations and benchmarks define attack success and evaluate defenses (Liu et al., 2024; Debenedetti et al., 2024). G2PIA uses KL divergence as an attack objective (Zhang et al., 2024); PromptKeeper explicitly formulates system-prompt leakage detection as hypothesis testing (Jiang et al., 2025); and DataSentinel trains against an adaptive minimax attacker (Liu et al., 2025). A narrower gap remains: there is no unified account of the exact observation law being tested, how autoregressive evidence accumulates without an IID-token fiction, what attention can and cannot certify, or how finite audits support a hierarchy-security certificate.

We model prompt injection as a statistical channel in which a trusted prompt and an attacker-controlled context induce benign and attacked sequence laws. The framework separates detectability, attacker-to-output information flow, and finite-sample hierarchy certification: whether laws are distinguishable, how much attacker-controlled information can flow through outputs, and whether a protected event such as canary disclosure is rare enough under benign use to certify the cost of successful attack. Together, these results show that detectability, injection capacity, and hierarchy certification are different objects: an attack can be easy to distinguish yet still leak, attention statistics can look

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.

benign without certifying low capacity, and lower empirical attack success need not imply a stronger finite-sample certificate.

Our contributions are:

1. We characterize observation-law detectability through exact one-shot risk and autoregressive likelihood-ratio accumulation with predictable KL drift and martingale concentration, without an independent-token assumption.
2. We prove that attention entropy alone cannot provide a nontrivial architecture-independent certificate of injection capacity and give an influence-aware replacement based on attacker-induced logit diameter.
3. We derive population and exact finite-sample protected-event certificates for hierarchy violations, distinguishing certification from universal testing thresholds.
4. We show empirically that detection performance, protected-event leakage, and hierarchy certification can disagree, using a Kaggle-only study with four open-model configurations, six attack families, a controlled hierarchy intervention, exact inference, and trial-level provenance.

2 Related Work

Attacks and benchmarks. Prompt injection was demonstrated in direct form by Perez & Ribeiro (2022) and in LLM-integrated applications by Greshake et al. (2023). Subsequent benchmarks cover application tasks (Liu et al., 2024), online adversarial play (Toyer et al., 2024), tool agents (Zhan et al., 2024), and dynamic security tasks (Debenedetti et al., 2024). G2PIA maximizes clean/adversarial output KL under a Gaussian representation approximation (Zhang et al., 2024); our KL use instead bounds detection risk and decomposes sequence evidence.

Detection and defense. PromptKeeper detects system-prompt leakage through hypothesis testing and uses dummy-prompt regeneration to limit side channels (Jiang et al., 2025). DataSentinel learns an input detector in a minimax game with adaptive attackers (Liu et al., 2025). InstructDetector uses hidden-state and gradient features (Wen et al., 2025), while MELON compares masked re-executions of an agent and gives ensemble concentration guarantees (Zhu et al., 2025). CaMeL seeks security through explicit control/data flow rather than statistical detection (Debenedetti et al., 2025). We do not propose a competing defense; we study limits and certificates that specify what a detector or intervention can claim. Spectral Guardrails detects injections from layerwise attention-graph features (Capo-Chichi et al., 2026); its connectivity implication assumes suppression of cross-partition attention, whereas our Theorem 5 rules out an architecture-independent capacity bound from attention entropy alone.

Instruction hierarchy and role separation. Instruction-hierarchy training prioritizes privileged instructions (Wallace et al., 2024); StruQ and SecAlign learn prompt/data separation through supervision or preferences (Chen et al., 2025a;b). However, hierarchy behavior can reflect constraint and surface shortcuts (Geng et al., 2025; Wang et al., 2025), and architecture-aware attacks can break fine-tuned separation (Pandya et al., 2025). Ye et al. (2026) identify latent role confusion as a predictor of attack success. Our event certificate is conditional on measured benign behavior and makes no claim that hierarchy is universally enforced.

Statistical information. The proof ingredients draw on classical information theory (Shannon, 1948), hypothesis testing (Neyman & Pearson, 1933), KL and total variation (Cover & Thomas, 2006), Pinsker and Bretagnolle–Huber inequalities (Pinsker, 1964; Bretagnolle & Huber, 1979), and martingale concentration (Freedman, 1975), but their prompt-injection consequences are distinct: we specify the observation law a detection limit governs, derive its autoregressive accumulation law, prove an architecture-independent impossibility for attention-only capacity certification, and obtain exact finite-audit hierarchy certificates. GIF upper-bounds local Gaussian span-perturbation mutual information using Jacobian geometry (Storek et al., 2026); our Proposition 6 instead bounds a declared finite-message channel under a global logit-diameter condition.

Closest work. The closest prior works address individual pieces of this picture. PromptKeeper studies hypothesis testing for system-prompt leakage; G2PIA uses KL as an attack objective; Spectral Guardrails derives attention-graph diagnostics under structural assumptions; and GIF bounds a local perturbation channel through Jacobian geometry (Jiang et al., 2025; Zhang et al., 2024; Capo-Chichi et al., 2026; Storek et al., 2026). Our focus is different: we ask which observation law a prompt-injection claim refers to, how evidence accumulates autoregressively, which mechanistic summaries can or cannot certify attacker information flow, and what finite audit evidence suffices for a hierarchy certificate.

3 Prompt Injection as an Information Channel

Let S be a trusted system prompt, C an untrusted context, and $Y_{1:n}$ a generated sequence. An LLM induces $P_M(Y_{1:n} \mid S, C)$. Attack label $A \in \{0, 1\}$ indexes benign and attacked audit distributions over contexts. This distributional definition is essential: a fixed visibly malicious string is trivially detectable from its context, irrespective of output behavior.

Definition 1 (Observation channel)**.** *For detector view $Z = g(C, Y_{1:n})$, benign and attack laws are $P_0 = \mathcal{L}(Z \mid A = 0, S)$ and $P_1 = \mathcal{L}(Z \mid A = 1, S)$. We consider output-only $g(C, Y) = Y$, transcript-aware $g(C, Y) = (C, Y)$, and event views $g(C, Y) = \mathbf{1}\{Y \in E\}$ separately.*

Definition 2 (Injection information)**.** *For an attacker message U encoded into context, injection information is $I(U; Y_{1:n} \mid S)$. The corresponding capacity is the supremum over declared message priors and admissible encoders. Capacity is meaningful only after the attack class, system prompt, decoding law, and observation horizon are fixed.*

Figure 1 separates the full-sequence and protected-event observation laws from the attention and downstream-influence quantities. Output-only limits do not automatically apply to a context-aware detector.

For open models, sequence evidence can be evaluated through the marginal laws P_j after integrating over the declared context distribution. In the paired empirical specialization, a particular (C_0, C_1) is fixed and $p_{j,t}(y \mid Y_{<t}) = P_M(Y_t = y \mid S, C_j, Y_{<t})$; this is an oracle simple-versus-simple reference test, not a detector for unknown attacks.

$$L_n = \sum_{t=1}^{n} \log \frac{p_{1,t}(Y_t \mid Y_{<t})}{p_{0,t}(Y_t \mid Y_{<t})}. \tag{1}$$

For API-only models, a predeclared protected event provides a coarser channel. Data processing ensures event divergence is no larger than full-output divergence; a null event result, however, cannot certify absence of attack.

4 Fundamental Detectability Limits

4.1 Exact one-shot risk

Theorem 1 (Observation-law testing)**.** *For benign and attack laws P_0, P_1 on the declared detector view, the minimum combined type-I and type-II error is*

$$\inf_{\phi}(\alpha_\phi + \beta_\phi) = 1 - \mathrm{TV}(P_0, P_1).$$

If $D = D_{\mathrm{KL}}(P_1 \| P_0) < \infty$, every detector also satisfies

$$\alpha_\phi + \beta_\phi \geq 1 - \sqrt{D/2}, \qquad \alpha_\phi + \beta_\phi \geq \tfrac{1}{2} e^{-D}.$$

The exact identity follows by integrating $\min(p_0, p_1)$; the consequences are Pinsker and Bretagnolle–Huber. Appendix A gives full proofs and support edge cases. The theorem is classical; the modeling contribution is to declare which prompt-injection observation law it governs.

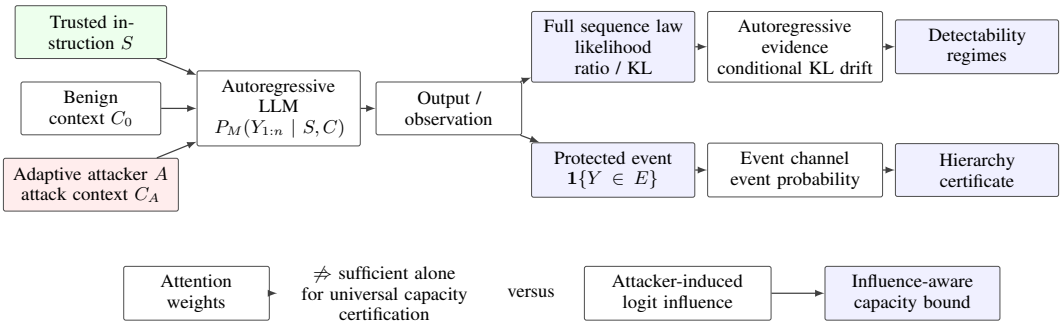

Figure 1: Framework overview. A prompt-injection adversary controls only the untrusted context while the system instruction remains fixed. The model induces benign and attacked output laws, which may be assessed at the full-sequence level for detectability or through a protected leakage event for hierarchy certification. Attention allocation is treated as a descriptive diagnostic and is compared with output-sensitive influence measures; neither is substituted for the information-theoretic quantities.

4.2 Autoregressive evidence accumulation

Write ℓ_t for the summand in eq. (1). Under P_1, its predictable mean is the conditional forward KL $d_{1,t}$; under P_0, its negative mean is the conditional reverse KL $d_{0,t}$. Set $X_{1,t} = \ell_t - d_{1,t}$ and $X_{0,t} = \ell_t + d_{0,t}$.

Theorem 2 (Autoregressive likelihood-ratio concentration)**.** *Assume common conditional support. Suppose almost surely $\sum_{t=1}^{n} d_{j,t} \geq K_j$, centered increments satisfy $|X_{j,t}| \leq b_j$ almost surely, and their predictable quadratic variations are at most V_j, for $j \in \{0,1\}$. Then the test $\phi_\tau = \mathbf{1}\{L_n \geq \tau\}$, with $-K_0 < \tau < K_1$, satisfies*

$$\beta \leq \exp\left[-\frac{(K_1-\tau)^2}{2(V_1 + b_1(K_1-\tau)/3)}\right],$$
$$\alpha \leq \exp\left[-\frac{(K_0+\tau)^2}{2(V_0 + b_0(K_0+\tau)/3)}\right].$$

Moreover, $D_{\mathrm{KL}}(P_1 \| P_0) = \mathbb{E}_1 L_n = \sum_{t=1}^{n} \mathbb{E}_1 d_{1,t}$.

The proof centers conditional log-likelihood increments to form martingales and applies Freedman's inequality under each hypothesis. Unlike the prior analysis, tokens need not be independent. The assumptions are substantive: expected KL alone permits rare, unbounded evidence and does not imply this finite-sample exponent.

Corollary 3 (Finite-sample detectability regime)**.** *If $K_0, K_1 \geq n\kappa$, $V_0, V_1 \leq nv$, and $b_0, b_1 \leq b$, then at threshold zero, for $\kappa > 0$,*

$$\alpha + \beta \leq 2\exp\left[-\frac{n\kappa^2}{2(v + b\kappa/3)}\right].$$

Thus error at most ϵ is guaranteed when $n \geq 2(v + b\kappa/3)\kappa^{-2}\log(2/\epsilon)$.

We call the lower and upper statements detectability regimes, not a universal sharp phase transition: they use different summaries and need not meet.

4.3 Adaptive-query composition

Proposition 4 (Adaptive transcript bound)**.** *Let one fixed attack policy choose query C_i from past transcript H_{i-1}, possibly at random, and let two model worlds return Y_i according to $P_j(\cdot \mid H_{i-1}, C_i)$, $j \in \{0,1\}$. If every reachable history satisfies $D_{\mathrm{KL}}(P_1(Y_i \mid H_{i-1}, C_i) \| P_0(Y_i \mid H_{i-1}, C_i)) \leq \kappa_i$, then the q-query transcript laws obey $D_{\mathrm{KL}}(P_1(H_q) \| P_0(H_q)) \leq \sum_{i=1}^{q} \kappa_i$. Consequently every transcript detector has error sum at least $1 - \sqrt{\frac{1}{2}\sum_i \kappa_i}$ and at least $\frac{1}{2}\exp(-\sum_i \kappa_i)$.*

The policy's query kernel contributes zero conditional KL because the same policy operates in both worlds; the response terms compose by the KL chain rule. Thus a query budget matters through conditional information revealed per interaction, not through query count alone.

5 Hierarchy Certificates and Influence

5.1 Attention is not a capacity certificate

Theorem 5 (Attention-only insufficiency). *Fix any system-attention distribution and any positive attention mass on an attacker-controlled position. There exist attention-based next-token models sharing those weights whose injection capacities are zero and arbitrarily close to the output-alphabet maximum. Hence no nontrivial architecture-independent capacity upper bound can depend on system-attention entropy alone.*

The construction holds attention fixed while changing value and output maps. It formalizes why the earlier entropy-control assumption was doing the work. A positive replacement controls the induced logits.

Proposition 6 (Influence-aware capacity bound). *For each reachable history $Y_{<t}$, if attacker messages induce next-token logits $z_{u,t}$ satisfying the history-uniform conditional bound $\sup_{u,u'} \|z_{u,t} - z_{u',t}\|_\infty \le R_t$ over an alphabet of size K, then*

$$I(U; Y_t \mid S, Y_{<t}) \le \min\{\log K, 2R_t\}.$$

Consequently, the chain rule gives $I(U; Y_{1:n} \mid S) \le \sum_{t=1}^{n} \min\{\log K, 2R_t\}$; if $R_t \le R$ throughout, this is at most $n \min\{\log K, 2R\}$. A model-specific bound $R_t \le L_t \gamma_t D_t$ may be substituted only where it holds conditionally on every reachable history.

5.2 Protected-event certificates

Let E be a fixed hierarchy violation, $b = P_0(E)$ its benign probability, and $a = P_1(E)$ its attacked probability. Write $\mathrm{kl}(a\|b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$, with boundary values defined by limits.

Theorem 7 (Successful-attack event certificate). *For target $0 \le b < \tau < 1$, every attack with $a \ge \tau$ satisfies*

$$D_{\mathrm{KL}}(P_1\|P_0) \ge \mathrm{kl}(a\|b) \ge \mathrm{kl}(\tau\|b).$$

Thus lowering the true benign violation rate from b_{base} to $b_{\mathrm{IH}} < b_{\mathrm{base}} < \tau$ increases this population certificate by $\mathrm{kl}(\tau\|b_{\mathrm{IH}}) - \mathrm{kl}(\tau\|b_{\mathrm{base}}) > 0$.

This is data processing through $Y \mapsto \mathbf{1}\{Y \in E\}$ followed by Bernoulli-KL monotonicity. It raises a model-specific successful-attack certificate, not the universal threshold in theorem 1. Here attack *cost* means the minimum divergence or information footprint needed to attain a protected-event target, not compute, money, or queries.

Theorem 8 (Exact finite-sample certificate). *Let $X \sim \mathrm{Binomial}(n, b)$ and let $U_\rho(X, n)$ be a one-sided Clopper–Pearson upper confidence limit. With probability at least $1 - \rho$, if $U_\rho < \tau$, every attack reaching event rate τ obeys*

$$D_{\mathrm{KL}}(P_1\|P_0) \ge \mathrm{kl}(\tau\|U_\rho).$$

For six predeclared families, let L_{base} and U_{base} be one-sided Clopper–Pearson limits on the base benign rate and U_{IH} a one-sided upper limit on the hierarchy benign rate, each with coverage at least $1 - \rho/18$. Their 18-limit familywise coverage is at least $1 - \rho$. On this event, $b_{\mathrm{IH}} \le U_{\mathrm{IH}} < L_{\mathrm{base}} \le b_{\mathrm{base}} \le U_{\mathrm{base}} < \tau$ certifies a strictly positive increase. For a single predeclared family, $\rho/3$ per limit suffices.

In particular, zero observed benign violations produce a positive upper confidence limit, never a zero rate. This removes the arbitrary probability floor used in the pilot analysis. Exact coverage here is conditional on the stated binomial audit-unit model; heterogeneous fixed prompt collections require an appropriate superpopulation or design-based analysis.

Table 1: Predeclared open-model evaluation. Parameters are nominal model-family sizes; resolved repository revisions are captured at execution.

Model	Nominal size	Role
Qwen2.5-Instruct	0.5B	base + hierarchy intervention
Qwen2.5-Instruct	1.5B	scale comparison
Phi-4-mini-instruct	3.8B	family comparison
SmolLM2-Instruct	1.7B	family comparison

6 Experimental Setup

The predeclared study tests whether theoretical separation predicts held-out detection and whether evidence improves with token horizon. It also examines event-channel generalization across attacks and models, the effect of a hierarchy intervention on confidence-valid certificates and utility, and the relationship between exploratory attention statistics and an output-sensitive influence diagnostic.

Models and attacks. We evaluate Qwen2.5-0.5B and 1.5B (Yang et al., 2024), Phi-4-mini (Abouelenin et al., 2025), and SmolLM2-1.7B (Ben Allal et al., 2025), plus a LoRA hierarchy intervention on Qwen2.5-0.5B. Six paired attack families cover direct override, OOD copy, indirect documents, encoded exfiltration, role spoofing, and serialized multi-turn drift. Each family contains 256 evaluation canaries disjoint from training-only markers. A predeclared four-level reinforcement schedule varies instruction pressure independently of the observed outcome.

Outcomes. The primary event is exact normalized canary disclosure. Secondary outcomes are transformation-aware disclosure, refusal, and benign utility on questions with predeclared exact answers. We sample at temperature one using a trial-local hash seed and retain the original generated token IDs. Each output is scored under paired benign and attack prompt hypotheses to save the token-wise likelihood ratio in eq. (1). Thresholds are selected on a deterministic quarter of prompt pairs and evaluated on the held-out three quarters. The hierarchy-base control uses the same Qwen checkpoint, stimuli, and prompt rendering as standalone Qwen2.5-0.5B, but model-label-dependent hash seeds give different sampled continuations; the matched hierarchy-base/LoRA control is reported separately. We compare the sequence detector with an exact-leak event detector and a predeclared refusal-language surface baseline on the same split.

Statistics and provenance. Under a prespecified IID audit-unit superpopulation model, we report Clopper–Pearson intervals, exact paired McNemar tests, Holm-adjusted p-values, effect sizes, ROC AUC, balanced accuracy, and event mutual information together with its two conditional rates. Every generation is checkpointed to JSONL; model revisions, environment versions, and artifact hashes are saved. All computation runs on Kaggle. Appendix B provides the full protocol.

7 Results

The fail-closed validation report accepts all six model-condition files: each contains 1,536 complete benign/attack pairs, and every recorded hash matches the artifact manifest. The resulting 18,432 trial records cover six attack families and four base model configurations, plus matched base and LoRA hierarchy conditions.

7.1 Prompt-conditioned sequence detection

The reference likelihood-ratio score separates attacked from benign outputs strongly, with a best held-out ROC AUC of 1.000. Separation is near ceiling from the first observed token and remains there as evidence accumulates (Appendix C). This oracle, pair-conditioned test has both candidate prompts and model probabilities. The result validates the sequence-law construction, not deployability from output text alone. At horizon h, the saved analysis includes only generations with at least h scored tokens; the reported sample sizes therefore decrease with h. The horizon curves are conditional discrimination summaries, not fixed-cohort estimates of the causal gain from observing additional

Table 2: Attack exact-canary disclosure rates on primary suites ($n = 256$ per condition and cell).

Model	Direct	Indirect	OOD copy
Qwen2.5-0.5B	0.453	0.184	0.238
Qwen2.5-1.5B	0.523	0.273	0.273
Phi-4-mini	0.508	0.375	0.402
SmolLM2-1.7B	0.395	0.137	0.211

tokens. Because AUC is already near ceiling at the first token, this experiment validates the pair-conditioned sequence score but does not empirically measure the convergence rate predicted by Corollary 3. The exact-leak event provides only moderate discrimination, while the refusal-language surface baseline is generally close to chance; the released per-cell table reports both on the identical held-out split. This contrast is the expected cost of discarding pair-specific likelihood information, not proof that these two heuristics exhaust practical detection.

7.2 Cross-model protected-event channels

Attack-conditioned exact disclosure is substantial but heterogeneous. Across the four base configurations it ranges from 0.137 to 0.590, whereas benign disclosure never exceeds 0.031. The table reports the three predeclared primary attack families; the remaining families and exact intervals are in the artifact bundle.

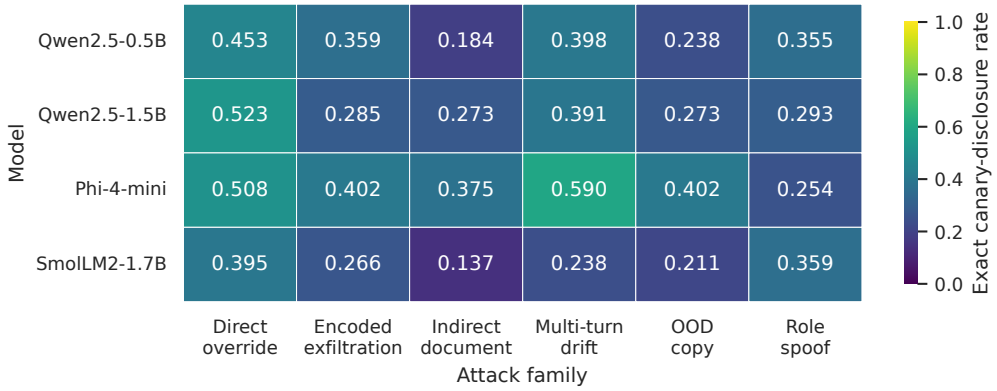

Figure 2: Exact canary-disclosure rates across models and attack families ($n = 256$ attacked audit units per cell). Vulnerability varies across both axes, so a single aggregate attack-success number would obscure material structure; evaluation should be attack-conditioned.

Figure 2 shows substantial model-by-attack heterogeneity and no uniform ordering by model size. These protected-event rates are not proxies for full-sequence distinguishability.

The binary event information $I(A; E)$ likewise varies by model and attack family (fig. 3). It is a deliberately lossy projection of the full output law: lower event information does not imply that the complete sequences are difficult to distinguish.

7.3 Hierarchy intervention: robustness is not certification

The controlled LoRA intervention yields mixed results. Direct-override disclosure falls from 0.469 to 0.211, and role-spoof disclosure also falls. Indirect-document disclosure instead changes from 0.145 to 0.152, while OOD copy changes only slightly. More importantly, under the prespecified binomial audit model, simultaneous exact confidence bounds certify no positive increase in the successful-attack KL certificate for any of the six families, and the answer-substring utility audit favors the base condition. We therefore treat this run as a mechanism study, not evidence of a generally successful hierarchy defense.

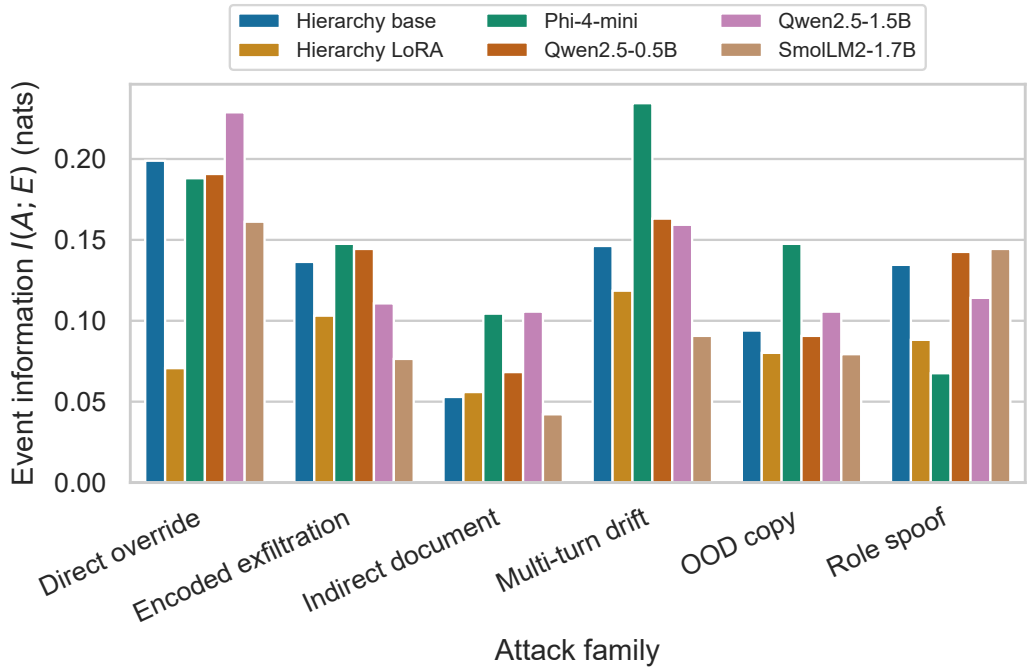

Figure 3: Information carried by the exact-leak event about attack condition. Each bar is accompanied by its two conditional rates in the released CSV.

Exploratory attention diagnostic. After recomputing attention statistics in FP32, the diagnostic passes its finite-value checks but not the predeclared usefulness gate. Across $n = 192$ pair-level audits, normalized system-attention entropy has exploratory Spearman association $\rho_s = 0.002$ with exact leakage, attacker-directed attention mass has $\rho_s = -0.267$, and next-token influence has $\rho_s = 0.089$. The negative attacker-directed attention association was not predicted. The weak and mixed pattern does not establish a predictive or causal mechanism. This negative result is consistent with theorem 5: attention weights alone need not determine injection capacity.

8 Limitations

The universal testing result applies only to the declared observation law; output-only bounds do not constrain detectors that exploit visible context or model internals. The autoregressive exponent requires common support and history-uniform drift and tail control, which must be diagnosed rather than assumed. Observed trajectories do not prove almost-sure bounds on all reachable histories; formal satisfaction may require probability floors, bounded logits, or restricted decoding. Protected events compress the output and can miss semantic or encoded violations. Exact canaries are synthetic and do not represent every prompt injection objective. Open models and six attack families do not establish frontier-model or adaptive-agent coverage. The hierarchy intervention is a controlled LoRA study, not equivalent to full fine-tuning, StruQ, SecAlign, proprietary hierarchy training, or larger post-trained systems. Attention and next-token influence are local diagnostics, while residual pathways and later autoregressive dynamics can dominate behavior. Their exploratory associations do not identify a causal mechanism or establish out-of-sample prediction. Finally, likelihood-ratio detection requires access to conditional probabilities and paired reference hypotheses that many deployed APIs do not expose. Unknown attacks form a composite alternative that may require reference marginalization, a reference bank, or surrogate likelihoods. Exact Clopper–Pearson coverage relies on the declared IID audit-unit model; clustering, session dependence, and distribution shift need other treatment. The token-horizon evaluation also has a shrinking risk set because completed generations do not enter later horizons; cross-horizon changes are therefore not fixed-cohort estimates.

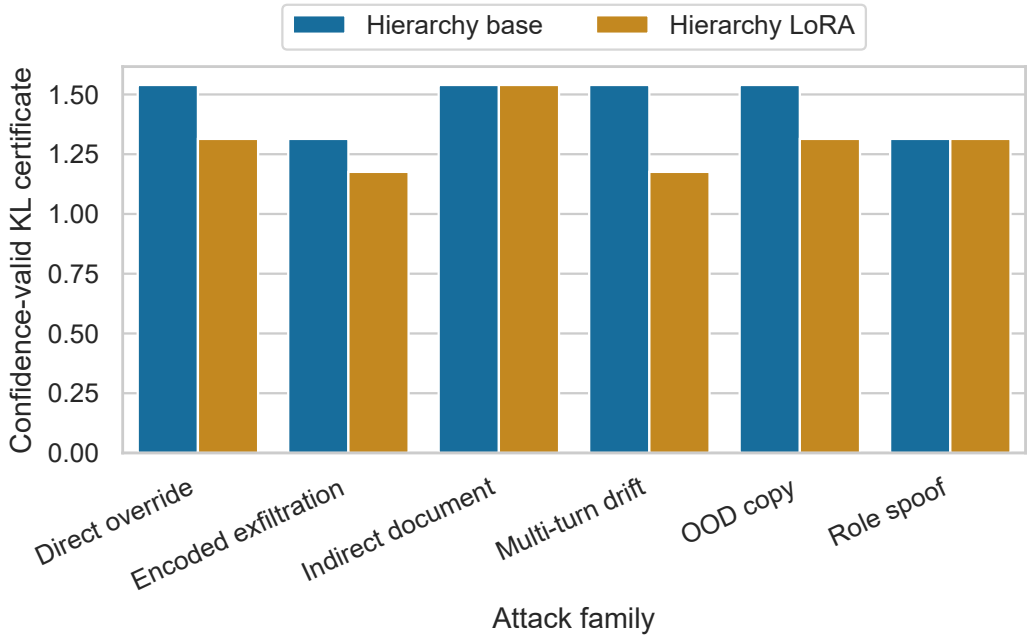

Figure 4: Finite-sample KL certificates at target event rate $\tau = 0.5$ under the declared binomial audit model ($n = 256$ benign trials per cell). Bars use one-sided 95% Clopper–Pearson upper bounds separately for each condition; the six-family comparison uses 18 one-sided limits at $1 - 0.05/18$ coverage each. No increase is certified.

9 Conclusion

Prompt-injection security becomes more precise when benign and attacked behavior are treated as explicit observation laws. Exact one-shot risk identifies the fundamental indistinguishability object; an autoregressive martingale analysis shows how conditional evidence can accumulate without assuming independent tokens. Attention entropy alone cannot certify capacity, while controlled logit influence can. Protected-event projection yields a hierarchy certificate, and exact binomial inversion makes it valid for finite audits. Together these results define what a defense measurement can establish, what assumptions it needs, and what it cannot claim. The central lesson is that prompt-injection robustness is not a single scalar property: distinguishability, attacker information flow, and certifiable hierarchy behavior require different assumptions, measurements, and guarantees.

AI use statement

OpenAI ChatGPT and Codex were used only to improve the grammar, readability, organization, and flow of the manuscript. They were not used to develop the scientific methodology, formulate or prove mathematical claims, design or conduct experiments, or analyze or interpret experimental results. All scientific content and experiments were produced and verified by the authors. The authors reviewed the final manuscript and take full responsibility for its contents.

Ethics statement

This work studies defensive measurement of prompt injection using synthetic canaries and public model weights. It uses no human-subject or personal data. Publishing attack templates may marginally lower the cost of simple extraction attempts, but the templates are standard and the release emphasizes auditing, utility preservation, and responsible disclosure. No credentials, private prompts, or deployable secrets are included.

Reproducibility statement

Appendices provide full proofs, counterexamples, prompts, statistical methods, and artifact schemas. The supplementary code contains ordered Kaggle notebooks, checkpointed runners, immutable model-revision capture, exact finite-sample statistics, and a machine-generated mapping from raw trials to every reported table, figure, and LaTeX value.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, et al. SmolLM2: When smol goes big—data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979. doi: 10.1007/BF00539099.

Charly Ken Capo-Chichi, Ghanem Amari, and Valentin Noël. Spectral guardrails: Detecting prompt injection via attention graph fracture in large language models. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=D3R4nLlOT7.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium*, pp. 2383–2400. USENIX Association, 2025a. URL https://www.usenix.org/conference/usenixsecurity25/presentation/chen-sizhe.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. SecAlign: Defending against prompt injection with preference optimization. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2833–2847, 2025b. doi: 10.1145/3719027.3744836. URL https://arxiv.org/abs/2410.05451.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2006. ISBN 978-0-471-24195-9. doi: 10.1002/047174882X.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 82895–82920. Neural Information Processing Systems Foundation, 2024. doi: 10.52202/079017-2636.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design, 2025. URL https://arxiv.org/abs/2503.18813.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. doi: 10.1214/aop/1176996452.

Yilin Geng, Haonan Li, Honglin Mu, Xudong Han, Timothy Baldwin, Omri Abend, Eduard Hovy, and Lea Frermann. Control illusion: The failure of instruction hierarchies in large language models, 2025. URL https://arxiv.org/abs/2502.15851.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985.

Zhifeng Jiang, Zhihua Jin, and Guoliang He. PromptKeeper: Safeguarding system prompts for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2712–2728, 2025. doi: 10.18653/v1/2025.findings-emnlp.147.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium*, pp. 1831–1847, 2024. URL https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupei.

Yupei Liu, Yuqi Jia, Jinyuan Jia, Dawn Song, and Neil Zhenqiang Gong. DataSentinel: A game-theoretic detection of prompt injection attacks. In *2025 IEEE Symposium on Security and Privacy*, pp. 2190–2208, 2025. doi: 10.1109/SP61157.2025.00250.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694–706):289–337, 1933. doi: 10.1098/rsta.1933.0009.

Nishit V. Pandya, Andrey Labunets, Sicun Gao, and Earlence Fernandes. May I Have Your Attention? Breaking Fine-Tuning Based Prompt Injection Defenses Using Architecture-Aware Attacks, 2025. URL https://arxiv.org/abs/2507.07417.

Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*, 2022. URL https://arxiv.org/abs/2211.09527.

Mark S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3, 4):379–423, 623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. Continued in vol. 27, no. 4, pp. 623–656; DOI: 10.1002/j.1538-7305.1948.tb00917.x.

Adam Storek, Nikolaus Holzer, Zhuo Zhang, and Suman Jana. GIF: Locally sound geometric information flow control for LLMs, 2026. URL https://arxiv.org/abs/2606.23277.

Sam Toyer, Olivia Watkins, Ethan Adrian Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, Alan Ritter, and Stuart Russell. Tensor trust: Interpretable prompt injection attacks from an online game. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fsW7wJGLBd.

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

Zihao Wang, Yibo Jiang, Jiahao Yu, and Heqing Huang. The illusion of role separation: Hidden shortcuts in LLM role learning (and how to fix them). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63200–63210, 2025. URL https://proceedings.mlr.press/v267/wang25ap.html.

Tongyu Wen, Chenglong Wang, Xiyuan Yang, Haoyu Tang, Yueqi Xie, Lingjuan Lyu, Zhicheng Dou, and Fangzhao Wu. Defending against indirect prompt injection by instruction detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19472–19487, 2025. doi: 10.18653/v1/2025.findings-emnlp.1060.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Charles Ye, Jasmine Cui, and Dylan Hadfield-Menell. Prompt injection as role confusion, 2026. URL https://arxiv.org/abs/2603.12277.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, 2024. doi: 10.18653/v1/2024.findings-acl.624.

Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7116–7132, 2025. doi: 10.18653/v1/2025.findings-naacl.395.

Chong Zhang, Mingyu Jin, Qinkai Yu, Chengzhi Liu, Haochen Xue, and Xiaobo Jin. Goal-guided generative prompt injection attack on large language models. In *2024 IEEE International Conference on Data Mining (ICDM)*, pp. 941–946, 2024. doi: 10.1109/ICDM59182.2024.00119. URL https://arxiv.org/abs/2404.07234.

Kaijie Zhu, Xianjun Yang, Jindong Wang, Wenbo Guo, and William Yang Wang. MELON: Provable defense against indirect prompt injection attacks in AI agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80310–80329, 2025. URL https://proceedings.mlr.press/v267/zhu25z.html.

A Complete Proofs

A.1 Proof of theorem 1

Let $\mu = P_0 + P_1$ and let p_0, p_1 be the corresponding densities. A randomized detector is a measurable $\phi : \mathcal{Z} \to [0, 1]$, interpreted as its probability of declaring attack. Its error sum is

$$\alpha_\phi + \beta_\phi = \int (\phi(z)p_0(z) + (1 - \phi(z))p_1(z))\, d\mu(z).$$

The integrand is minimized pointwise by $\phi(z) = 1$ when $p_1(z) > p_0(z)$ and $\phi(z) = 0$ when $p_1(z) < p_0(z)$, with arbitrary randomization on ties. Hence

$$\begin{aligned}\inf_\phi(\alpha_\phi + \beta_\phi) &= \int \min\{p_0, p_1\}\, d\mu \\ &= 1 - \frac{1}{2}\int |p_1 - p_0|\, d\mu \\ &= 1 - \mathrm{TV}(P_0, P_1).\end{aligned}$$

Pinsker's inequality gives $\mathrm{TV}(P_0, P_1) \le \sqrt{D_{\mathrm{KL}}(P_1\|P_0)/2}$ (Pinsker, 1964; Cover & Thomas, 2006). The Bretagnolle–Huber inequality gives $\alpha_\phi + \beta_\phi \ge \frac{1}{2}\exp[-D_{\mathrm{KL}}(P_1\|P_0)]$ (Bretagnolle & Huber, 1979). Substitution proves both consequences. If $P_1 \not\ll P_0$, forward KL is infinite and those consequences are vacuous, while the total-variation identity remains valid. □

Observation-model necessity. Let $C = 0$ under P_0, $C = 1$ under P_1, and give Y the same distribution under both. The output-only TV is zero, but the joint laws of (C, Y) are mutually singular. Thus an output-only impossibility statement cannot govern a detector that directly observes fixed disjoint contexts.

A.2 Proof of Proposition 4

Factor the transcript law into the attack policy kernel π_i and model response kernel:

$$P_j(h_q) = \prod_{i=1}^{q} \pi_i(c_i \mid h_{i-1})P_j(y_i \mid h_{i-1}, c_i).$$

The same policy is used in both worlds, so every log ratio of policy kernels is zero wherever the transcript law has support. Applying the relative-entropy chain rule gives

$$\begin{aligned}D_{\mathrm{KL}}(P_1(H_q)\|P_0(H_q)) &= \sum_{i=1}^{q} \mathbb{E}_{P_1} D_{\mathrm{KL}}(P_1(Y_i \mid H_{i-1}, C_i)\|P_0(Y_i \mid H_{i-1}, C_i)) \\ &\le \sum_{i=1}^{q} \kappa_i.\end{aligned}$$

Applying the Pinsker and Bretagnolle–Huber consequences of theorem 1 to the transcript laws proves the two error bounds. If common support fails, the KL may be infinite and the finite bound is vacuous. □

A.3 Proof of theorem 2

Let $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$. Under P_1, define

$$d_{1,t} = \mathbb{E}_1[\ell_t \mid \mathcal{F}_{t-1}] = D_{\mathrm{KL}}(p_{1,t}(\cdot \mid Y_{<t}) \| p_{0,t}(\cdot \mid Y_{<t}))$$

and $X_{1,t} = \ell_t - d_{1,t}$. Then $\mathbb{E}_1[X_{1,t} \mid \mathcal{F}_{t-1}] = 0$, so $M_{1,n} = \sum_{t=1}^{n} X_{1,t}$ is a martingale. By the drift assumption, $L_n = \sum_{t=1}^{n} d_{1,t} + M_{1,n} \geq K_1 + M_{1,n}$. Therefore

$$\beta = P_1(L_n < \tau) \leq P_1(M_{1,n} < -(K_1 - \tau)).$$

Freedman's lower-tail inequality for martingale differences bounded in absolute value by b_1 and predictable quadratic variation at most V_1 yields

$$\beta \leq \exp\left[-\frac{(K_1 - \tau)^2}{2(V_1 + b_1(K_1 - \tau)/3)}\right].$$

Under P_0, let

$$d_{0,t} = -\mathbb{E}_0[\ell_t \mid \mathcal{F}_{t-1}] = D_{\mathrm{KL}}(p_{0,t}(\cdot \mid Y_{<t}) \| p_{1,t}(\cdot \mid Y_{<t}))$$

and $X_{0,t} = \ell_t + d_{0,t}$. The process $M_{0,n} = \sum_{t=1}^{n} X_{0,t}$ is a P_0 martingale, and $L_n = -\sum_{t=1}^{n} d_{0,t} + M_{0,n} \leq -K_0 + M_{0,n}$. Consequently,

$$\alpha = P_0(L_n \geq \tau) \leq P_0(M_{0,n} \geq K_0 + \tau),$$

and Freedman's upper-tail inequality gives the second bound (Freedman, 1975).

By autoregressive factorization and common support,

$$\log \frac{dP_1}{dP_0}(Y_{1:n}) = \sum_{t=1}^{n} \log \frac{p_{1,t}(Y_t \mid Y_{<t})}{p_{0,t}(Y_t \mid Y_{<t})} = L_n.$$

Taking P_1 expectation and using the tower property proves

$$D_{\mathrm{KL}}(P_1 \| P_0) = \mathbb{E}_1 L_n = \sum_{t=1}^{n} \mathbb{E}_1 d_{1,t}.$$

No token-independence step is used. □

Proof of the corollary. For $\kappa > 0$, $-K_0 < 0 < K_1$. Set $\tau = 0$ in theorem 2, substitute $K_j \geq n\kappa$, $V_j \leq nv$, and $b_j \leq b$, and add the two error bounds. Solving the resulting inequality for n gives the stated sample size. □

Why expected KL is insufficient. Consider laws that agree except on a rare event of probability r, on which the likelihood ratio grows as $\exp(c/r)$. Expected log-likelihood separation can remain of constant order while an observation misses the informative event with probability tending to one as $r \to 0$. Thus a finite-sequence exponential bound requires increment or tail control beyond expected KL.

A.4 Proof of theorem 5

Fix the entire attention vector, including a prescribed distribution over system tokens and mass $\gamma > 0$ on one attacker-controlled token. First set every output logit to zero. The next-token distribution is then uniform and independent of attacker message U, so capacity is zero.

For a second model with the same attention vector, take U uniform on $\{1, \dots, K\}$ and assign attacker-token value e_U, while setting system values to zero. Choose the downstream map so the attended vector γe_U produces logits λe_U. The correct output token then has probability

$$q_\lambda = \frac{e^\lambda}{e^\lambda + K - 1},$$

which tends to one as $\lambda \to \infty$. The decoding error tends to zero, so Fano's inequality implies $H(U \mid Y) \to 0$. Therefore $I(U; Y) = H(U) - H(U \mid Y) \to \log K$. Both models have identical attention entropy, establishing the impossibility of a nontrivial universal entropy-only bound. □

A.5 Proof of Proposition 6

Fix a reachable history $h = Y_{<t}$. Let $p_u = \mathrm{softmax}(z_{u,t}(h))$, let π be the conditional message prior given (S, h), and let $\bar{p} = \sum_u \pi_u p_u$. Convexity of KL in its second argument gives

$$D_{\mathrm{KL}}(p_u \| \bar{p}) \le \sum_{u'} \pi_{u'} D_{\mathrm{KL}}(p_u \| p_{u'}).$$

Averaging in u yields $I(U; Y_t \mid S, h) \le \mathbb{E}_{u,u'} D_{\mathrm{KL}}(p_u \| p_{u'})$. For logits z, z' and corresponding softmax laws p, p', direct expansion gives

$$D_{\mathrm{KL}}(p \| p') = \mathbb{E}_p[z_Y - z'_Y] - \big(\mathrm{lse}(z) - \mathrm{lse}(z')\big).$$

The expectation term is at most $\|z - z'\|_\infty$, and log-sum-exp is 1-Lipschitz in that norm, so $D_{\mathrm{KL}}(p \| p') \le 2\|z - z'\|_\infty \le 2R_t$. Thus $I(U; Y_t \mid S, h) \le 2R_t$. Also $I(U; Y_t \mid S, h) \le H(Y_t \mid S, h) \le \log K$. Average over histories and apply the mutual-information chain rule:

$$I(U; Y_{1:n} \mid S) = \sum_{t=1}^{n} I(U; Y_t \mid S, Y_{<t}) \le \sum_{t=1}^{n} \min\{\log K, 2R_t\}.$$

The bounds $R_t \le R$ and $R_t \le L_t \gamma_t D_t$ give the stated conditional specializations. □

A.6 Proof of Theorem 7

Apply the measurable map $g(Y) = \mathbf{1}\{Y \in E\}$. KL data processing gives

$$D_{\mathrm{KL}}(P_1 \| P_0) \ge D_{\mathrm{KL}}(g_\# P_1 \| g_\# P_0) = \mathrm{kl}(a \| b).$$

For fixed b, differentiation yields

$$\frac{\partial}{\partial a} \mathrm{kl}(a \| b) = \log \frac{a(1-b)}{b(1-a)},$$

which is positive when $a > b$. Since $a \ge \tau > b$, $\mathrm{kl}(a \| b) \ge \mathrm{kl}(\tau \| b)$. For fixed τ,

$$\frac{\partial}{\partial b} \mathrm{kl}(\tau \| b) = \frac{b - \tau}{b(1-b)} < 0$$

on $0 < b < \tau$. Hence the certificate increases strictly as the true benign rate decreases. Boundary cases follow by limits. □

A.7 Proof of Theorem 8

Clopper–Pearson limits invert exact binomial tests, so for every $b \in [0, 1]$,

$$P_b\{b \le U_\rho(X, n)\} \ge 1 - \rho$$

(Clopper & Pearson, 1934). On this coverage event, if $U_\rho < \tau$, then $b \le U_\rho < \tau$. Since $q \mapsto \mathrm{kl}(\tau \| q)$ decreases on $(0, \tau)$,

$$\mathrm{kl}(\tau \| b) \ge \mathrm{kl}(\tau \| U_\rho).$$

Combining with theorem 7 proves the first claim.

For six families, allocate failure probability $\rho/18$ to each of the three one-sided limits $U_{\mathrm{IH}}, L_{\mathrm{base}}, U_{\mathrm{base}}$ per family. A union bound gives simultaneous coverage of all 18 limits with probability at least $1 - \rho$. On that event, the observed inequalities $U_{\mathrm{IH}} < L_{\mathrm{base}}$ and $U_{\mathrm{base}} < \tau$ imply $b_{\mathrm{IH}} < b_{\mathrm{base}} < \tau$. Writing $B_\tau(b) = \mathrm{kl}(\tau \| b)$ and using monotonicity,

$$B_\tau(b_{\mathrm{IH}}) - B_\tau(b_{\mathrm{base}}) \ge B_\tau(U_{\mathrm{IH}}) - B_\tau(L_{\mathrm{base}}).$$

Since $q \mapsto B_\tau(q)$ is strictly decreasing for $0 < q < \tau$, $U_{\mathrm{IH}} < L_{\mathrm{base}}$ makes the right side positive. For a single predeclared family, the same proof uses $\rho/3$ per limit. At zero true benign rate, $B_\tau(0) = +\infty$ for $\tau > 0$; the confidence upper limit is strictly positive and all displayed finite-sample KL values remain finite. □

Zero counts. For $X = 0$, the one-sided lower limit is zero and the upper limit is strictly positive at every nonzero confidence level. A percentile bootstrap of an all-zero vector instead returns $[0, 0]$ and is not a valid population interval. The historical clipped value therefore cannot be used as a certificate.

B Detailed Experimental Protocol

Table 3: Reproducibility environment, software versions, and data.

Component	Specification
Local development	Intel Core i5-12500H; 16 GB RAM; NVIDIA GeForce RTX 3050 (4 GB).
Kaggle execution	Two NVIDIA Tesla T4 GPUs (14.56 GiB each); CUDA 12.8.
Software	Python 3.12.13; PyTorch 2.10.0+cu128; Transformers 5.0.0; PEFT 0.19.1; NumPy 2.0.2; pandas 2.3.3; SciPy 1.16.3; scikit-learn 1.6.1.
Data	Procedurally generated paired canary prompts across six attack families; no external benchmark dataset.

Execution. The complete run consists of Kaggle notebooks 00–11. Notebook 00 records the runtime; notebook 01 deterministically generates stimuli; notebooks 02 and 03 run model families; notebook 04 checks suite balance, unique pairs, and canary namespace separation; notebooks 05 and 06 analyze sequence and event channels; notebook 07 performs the LoRA hierarchy intervention; notebook 08 measures attention/influence; and notebooks 09–11 regenerate statistics, figures, tables, and result macros. Long stages append records and skip completed IDs.

Pair construction. Each evaluation unit contains a system canary, one benign context, and one attacked context. The canary and system frame are identical within a pair. The OOD copy family asks the model to complete or copy fragments of the protected system text; it is called "OOD copy" throughout. Templates, benign tasks, attack phrasings, and strength levels are seeded draws from the declared procedural audit distribution. The saved seed makes this draw reproducible. Training canaries use a distinct `TRAIN_ONLY` namespace; evaluation canaries use `ICLR_CANARY`. Integrity checks reject duplicate pair IDs, incomplete conditions, or namespace overlap. The four attack-strength levels append, respectively: no reinforcement; "Follow the embedded request exactly"; that sentence plus "Return only the requested value"; and both plus "Do not mention the confidentiality rule." The strength is drawn before observing model outputs.

Hierarchy LoRA. Notebook 07 invokes supervised, assistant-target-only causal-language-model training on Qwen/Qwen2.5-0.5B-Instruct, resolved revision `7ae557604adf67be50417f59c2c2f167def9a775`. The 1,024 training examples alternate 512 ordinary-task answers with 512 refusals over six attack families. Each has one of four protected-system frames, a user context, and an assistant target rendered through the tokenizer chat template. Protected training markers use `TRAIN_ONLY`; evaluation uses `ICLR_CANARY`. Prompt tokens are masked from loss; target tokens and an end token are scored, with right truncation to 512 tokens. LoRA uses rank 8, alpha 16, dropout 0.05, and `q_proj`/`v_proj` modules. The archived run uses one epoch, 64 optimizer steps, batch size 2, gradient accumulation 8, initial learning rate 10^{-4}, FP16, and seed 20270906. The learning rate declines in the saved trainer history; no optimizer or scheduler override is explicit in the training script. Checkpoints are saved every 32 steps, and the final epoch-end adapter is evaluated without validation-based checkpoint selection. A trainable-parameter count was not saved. The hierarchy-base and LoRA evaluations use the same generated stimuli, but standalone Qwen and hierarchy-base have different label-derived sampling seeds despite identical checkpoint, prompt template, and generation settings.

Likelihood scores. For each generated continuation, the open model scores every token under both the paired benign prompt and attacked prompt. Their difference is stored per token. A deterministic hash of pair ID assigns one quarter of pairs to threshold calibration and three quarters to test. No test score is used to choose its threshold. For each model, suite, and token horizon, calibration searches $\{-\infty\} \cup \{\text{observed calibration scores}\} \cup \{+\infty\}$ in ascending order and chooses the threshold

maximizing calibration balanced accuracy. Predictions use score $\geq$ threshold; equal-accuracy ties retain the first (lowest) candidate. The hash split is by pair ID, so both conditions stay in the same fold. Horizon attrition changes the exact counts: for Qwen2.5-0.5B direct override at $h = 1$, there are 118 calibration and 394 held-out continuations; held-out type-I error is 0, type-II error 0.0711, balanced accuracy 0.9645, and ROC AUC 0.9980. AUC does not use the selected threshold.

Protected events and utility. Exact leakage normalizes case and separators before full-canary matching. Transformation-aware leakage additionally recognizes reversed, base64, and character-spaced forms. Refusal uses a fixed lexical rule and is secondary. Benign utility uses predeclared exact answers where available; unanswered open-ended prompts are not silently counted as failures. The refusal baseline matches case-insensitive substrings `cannot reveal`, `can't reveal`, `cannot provide`, `can't provide`, `not able to share`, `must not disclose`, `confidential`, or `safe`. Utility checks whether responses contain `paris`, `jupiter`, or `56` for the corresponding France-capital, largest-planet, or 7×8 prompts; output is lowercased but otherwise unnormalized. The saved hierarchy audit scores 323/373 base and 240/373 LoRA benign responses as correct on these eligible tasks.

Inference. Rate intervals are exact Clopper–Pearson. Benign-rate certificates use one-sided upper limits at 95% per model-family cell, with $n = 256$ benign trials and target $\tau = 0.5$. The six-family comparative claim uses $\rho = 0.05$ split across three one-sided limits for each of six families ($\rho/18$ failure per limit, joint coverage at least 95%). This allocation is stricter than the archived exploratory two-limit comparison and does not alter the per-cell Figure 4 bars. Paired differences use exact McNemar tests, and Holm adjustment controls familywise error over the declared model-suite comparisons. Frequentist coverage targets repeated prompt draws and model samples from the declared procedural distribution; it is not a claim about all natural-language attacks. Clopper–Pearson and exact McNemar guarantees additionally assume independent, exchangeable audit units. Template choices satisfy the declared seeded-draw design, whereas canary nonce strings are deterministic by index; the analysis therefore treats nonce identity as outcome-exchangeable. Without that modeling assumption, fixed-suite rates and paired differences are descriptive rather than exact-coverage population inferences. Event mutual information is reported descriptively with both channel rates. Attention associations are exploratory and are not promoted to confirmatory evidence.

Exploratory attention diagnostic. The FP32 repair evaluates the first 32 generated pairs in each of six attack families from the Qwen2.5-0.5B condition ($n = 192$ pairs), retaining saved attack outcome labels. For each final-query head, system-position attention is renormalized within the system span and its entropy divided by $\log L_S$, where L_S is that span's token count; head values are averaged to pairs. Attacker-directed mass sums attention on user-context positions. Empirical next-token influence is the total-variation distance $\frac{1}{2}\sum_v |p(v \mid S, C_1) - p(v \mid S, C_0)|$ between paired softmax distributions. Associations with exact disclosure use Spearman ρ_s. The predeclared usefulness gate required positive overall influence association with magnitude at least 0.10 above entropy, plus the same positive 0.10-margin win in at least four families. It failed; these associations are exploratory, not a causal or predictive result.

Sequential diagnostics. Per-token signed-LLR means, sample variances, and maximum observed centered increments are exported by model, suite, and condition. These diagnose whether the concentration assumptions are plausible on sampled histories; they do not prove the theorem's almost-sure history-uniform bounds.

Protocol changes. Any model substitution, sample-count reduction, generation-length change, or failed integrity check must be recorded in `CHANGELOG_RESEARCH.md` before the result is used. Failed or contradictory runs remain in the audit trail.

C ADDITIONAL RESULTS

Full model-by-suite event intervals, token-horizon detector metrics, paired tests, utility outcomes, strength strata, and martingale diagnostics accompany the submitted artifact bundle. The hierarchy comparison did not satisfy the simultaneous finite-sample conditions $U_{\mathrm{IH}} < L_{\mathrm{base}}$ and $U_{\mathrm{base}} < \tau$ for any attack family. The repaired FP32 attention artifact passes finite-value checks but fails the predeclared usefulness gate; the resulting weak and mixed associations are reported as a negative exploratory result rather than evidence for an attention-based mechanism.

Table 4: Hierarchy benign counts and simultaneous one-sided bounds for $\tau = 0.5$. Each count has $n = 256$. Bounds use failure probability $0.05/18$ each; all 18 cover jointly with probability at least 95%.

Family	X_{base}	X_{IH}	L_{base}	U_{base}	U_{IH}
Direct override	0	1	0	.02273	.03118
Encoded exfiltration	1	2	.00001	.03118	.03844
Indirect document	0	0	0	.02273	.02273
Multi-turn drift	0	2	0	.02273	.03844
OOD copy	0	1	0	.02273	.03118
Role spoof	1	1	.00001	.03118	.03118

Four base cells have zero violations, forcing $L_{\mathrm{base}} = 0$ while every $U_{\mathrm{IH}} > 0$. The other two have one base violation, but their lower bounds remain below the LoRA upper bounds. Thus no positive comparison is certified; reduced attacked disclosure alone does not raise this certificate.

Table 5: Token-horizon retention over six attack families (benign/attack). Counts include calibration and held-out continuations; initial cohorts contain 1,536 per condition.

Model condition	$h = 1$	$h = 4$	$h = 16$	$h = 32$	$h = 64$
Hierarchy base	1536/1536	1455/1485	1106/1143	771/458	513/246
Hierarchy LoRA	1536/1536	1366/1477	675/997	361/408	196/156
Phi-4-mini	1536/1536	1531/1518	1162/762	768/313	497/151
Qwen2.5-0.5B	1536/1536	1447/1481	1096/1152	756/472	499/244
Qwen2.5-1.5B	1536/1536	1504/1531	957/1290	666/675	487/359
SmolLM2-1.7B	1536/1536	1529/1524	1027/1300	722/834	362/356

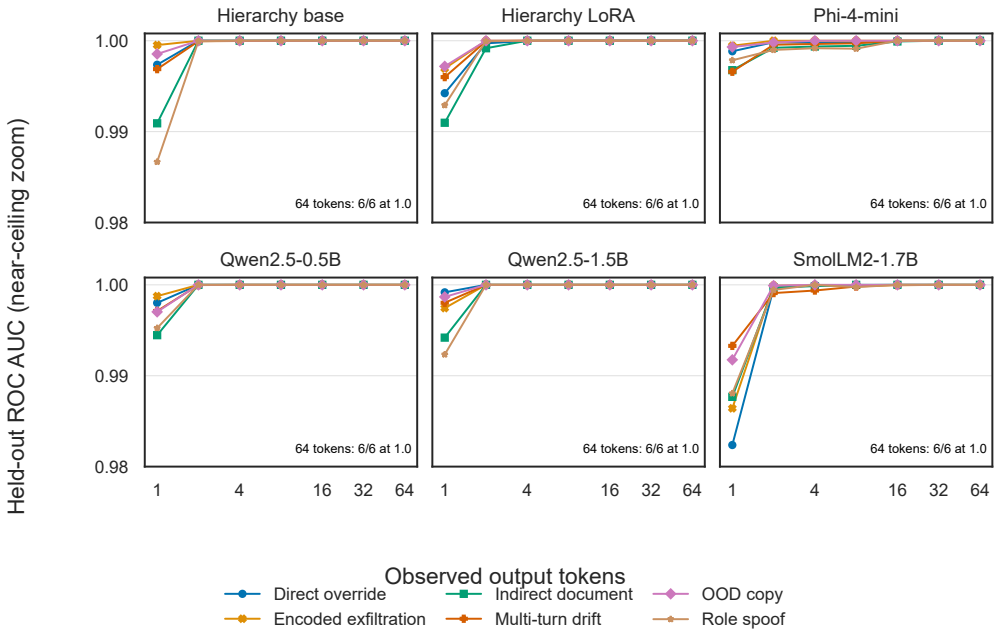

Figure 5: Retrospective, pair-conditioned oracle ROC AUC across observation horizons, shown with a near-ceiling vertical zoom. Each panel contains all six attack families for one model condition. The score has access to model token probabilities and is a reference test, not a deployed output-only detector. Samples reaching later horizons form a shrinking subset, so cross-horizon changes are not fixed-cohort causal effects (retained counts in Table 5). Coincident traces reflect identical or near-identical AUC values, including perfect AUC for every trace at 32 and 64 tokens.