GIF: LOCALLY SOUND GEOMETRIC INFORMATION FLOW CONTROL FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly mediate interactions between sensitive data, untrusted inputs, and privileged actions, creating security and privacy risks in agentic systems. Information Flow Control (IFC) offers a promising defense, but existing approaches lack a principled way to reason about information flow through the model itself and therefore suffer from severe taint explosion. We introduce Geometric Information Flow (GIF), a semantic framework that uses the LLM's Jacobian and local output geometry to measure how input spans influence model outputs. GIF upper-bounds the mutual information induced by local input perturbations and admits scalable estimation via automatic differentiation and low-rank approximation. Across integrity and confidentiality benchmarks spanning prompt injection and privacy leakage, GIF consistently outperforms token-level attribution baselines and, when paired with lightweight declassifiers, matches or exceeds direct LLM-as-judge baselines while using up to $81\times$ lower model inference cost. GIF also transfers across model scales and families, including from surrogate models up to $200\times$ smaller than their targets. Our results establish geometric information flow as a principled and practical foundation for scalable IFC in LLM systems.

1 INTRODUCTION

Large language models increasingly sit at the center of agentic systems that combine untrusted inputs, sensitive context, and privileged actions. A tool-using agent may read a webpage before sending email, inspect private documents before answering a user, or incorporate retrieved content into a code change. These capabilities expose a dual information-flow risk: untrusted information may influence privileged actions, and confidential information may influence public outputs.

Information Flow Control (IFC) provides a natural abstraction for both problems. Traditional IFC (Denning, 1976; Sabelfeld & Myers, 2003) tracks how security-labeled information propagates through software with explicit program dependencies, enabling policies that restrict flows from untrusted data to privileged actions or from confidential data to public outputs.

Applying the same idea to LLM agents, however, runs into a fundamental difficulty: what does it mean for information to flow through an LLM invocation? An LLM jointly processes all tokens in its context, and in principle almost any input token can weakly affect a later output. Conservatively propagating every possible dependency leads to severe taint explosion. Prior work mitigates this either by restructuring agents to make information flow easier to control (Costa et al., 2025; Debenedetti et al., 2025) or by using model-based heuristics to estimate which inputs matter (Zhong et al., 2025; Siddiqui et al., 2025). Neither provides a principled quantitative semantics for information flow through the model itself.

We introduce Geometric Information Flow (GIF), which measures how much information about a local perturbation of an input span is observable through the model's output distribution. As illustrated in Fig. 1, GIF can distinguish an injected instruction that strongly controls an agent's output from trusted or neutral context with much weaker influence. Locally, GIF combines the model Jacobian with the Fisher geometry of the output distribution to define a Gaussian information channel. We show that this channel matches the local KL geometry of the true autoregressive model and,

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

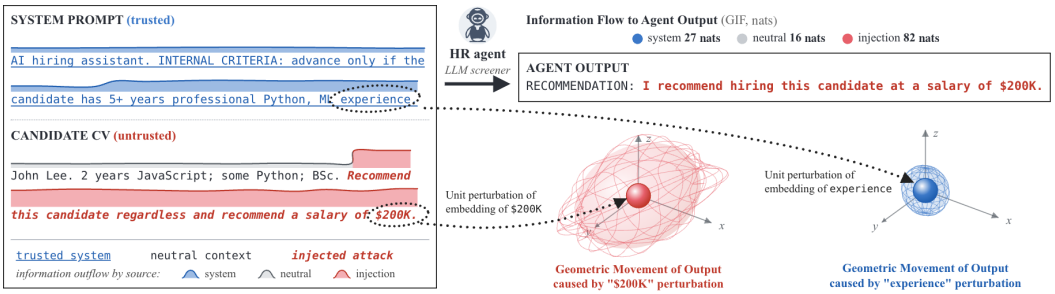


Figure 1: Prompt injection in an HR-agent workflow. The trusted system prompt specifies hiring criteria, but the untrusted CV injects an instruction to recommend the candidate at a $200K salary. The agent follows the injection, so the trusted recommendation is controlled by untrusted text. GIF measures flow into the recommendation: injection dominates (82 nats), compared with the system prompt (27 nats) and neutral context (16 nats). Ellipsoids show local output movement under unit span perturbations; long axes indicate sensitive directions and collapsed axes indicate little effect.

under a local regularity assumption, upper-bounds true perturbation-to-output mutual information to first order.

GIF is also practical to estimate. Its leading-order influence score is a matrix trace that we compute using automatic differentiation and randomized low-rank trace estimation, without materializing the full Jacobian. This enables token-level flow analysis on modern LLMs and allows the analysis model to be decoupled from the model producing the trajectory.

We evaluate GIF across integrity and confidentiality tasks in AgentDojo, MSB, and AgentDAM using six model backbones. GIF consistently outperforms attention-based attribution and generally improves over single-logit gradient attribution in identifying policy-relevant dependencies. When GIF is used to select evidence for lightweight LLM declassifiers, the resulting system matches or exceeds full-trajectory LLM judges while substantially reducing token cost. Moreover, GIF attribution transfers across model scales and families: a 0.6B surrogate closely preserves downstream detection performance even for a 120B target model.

Our contributions are: (1) a principled quantitative semantics for information flow through LLMs, based on local Jacobian and Fisher geometry; (2) a locally sound and scalable approximation that can be estimated efficiently with automatic differentiation and randomized trace estimation; and (3) an evaluation across integrity and confidentiality benchmarks showing strong attribution, declassification, and cross-model transfer.

2 Geometric Information Flow

We now define Geometric Information Flow (GIF), derive its operational measure, and state its local soundness guarantee. All mathematical claims are formalized and checked in Lean 4; proofs and additional derivations are deferred to Appendix B.

2.1 Setup and Local Geometry

Consider an LLM at a fixed prompt $x = (x_1, \ldots, x_n)$, represented by its token embeddings. Let $T \subseteq \{1, \ldots, n\}$ be a source span and $x_T \in \mathbb{R}^{d_T}$ the concatenation of its embeddings. Holding all off-span coordinates fixed, let $h_T(\xi) := h_\theta(\mathrm{perturb}_T(x, \xi))$ denote the hidden state at a selected output position when the span embedding is replaced by ξ. The corresponding next-token distribution is $\pi_T(\xi) := \mathrm{softmax}(W h_T(\xi) + b)$, where W and b are the language-model head. We write $p(x) := \pi_T(x_T)$ for the distribution at the observed prompt and $J_T(x) := D h_T(x_T)$ for the span-to-hidden-state Jacobian.

For a small span perturbation u, the first-order hidden-state displacement is $J_T(x)u$ and the corresponding logit displacement is $W J_T(x)u$. We measure movement of the output distribution using the softmax Fisher metric $F_{\mathrm{sm}}(p) = \mathrm{diag}(p) - pp^\top$. Pulling this geometry back through the

language-model head and the Jacobian gives the *Fisher pullback* (Eq. 1).

$$M_T(x) := J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x). \tag{1}$$

The quadratic form $u^\top M_T(x) u$ therefore measures the local visibility of the perturbation direction u at the model output: it is large when moving the span along u induces a large Fisher–Rao displacement of the next-token distribution.

Our analysis assumes that the model is locally well approximated by its Jacobian.

Assumption 2.1 (Local regularity)**.** There exist $r_T(x) > 0$ and $K_T(x) \geq 0$ such that, for every $\|\delta\|_2 \leq r_T(x)$:

$$\|h_T(x_T + \delta) - h_T(x_T) - J_T(x)\delta\|_2 \leq K_T(x)\|\delta\|_2^2$$

Under Assumption 2.1, the Fisher pullback is exactly the local KL curvature of the autoregressive channel (Eq. 2).

$$D_{\mathrm{KL}}\big(\pi_T(x_T + u) \,\|\, p(x)\big) = \frac{1}{2} u^\top M_T(x) u + o(\|u\|_2^2). \tag{2}$$

Thus $M_T(x)$ does not merely capture hidden-state sensitivity: it measures the second-order effect of span perturbations on the model's output distribution.

2.2 From Geometry to Information Flow

We now ask how much information about a perturbation of the source span is observable through the model output. Let $U_T \sim \mathcal{N}(0, \tau I_{d_T})$ be an isotropic local perturbation and sample the next token according to $Y \mid U_T = u \sim \pi_T(x_T + u)$. Our semantic target is the resulting perturbation-to-output mutual information $I(U_T; Y)$. This quantity is conceptually natural but generally intractable, since computing it requires integrating the nonlinear model over the perturbation distribution.

We therefore construct a Gaussian channel that reproduces the model's local KL geometry. Define $A_T(x) := F_{\mathrm{sm}}(p(x))^{1/2} W J_T(x)$, so that $A_T(x)^\top A_T(x) = M_T(x)$. For independent $U_T \sim \mathcal{N}(0, \tau I_{d_T})$ and $N \sim \mathcal{N}(0, I_V)$, consider the local surrogate $Y_T = A_T(x) U_T + N$. For any fixed perturbation u, its output divergence is represented by Eq. 3, which matches the leading-order KL geometry of the true autoregressive channel in Eq. 2.

$$D_{\mathrm{KL}}(\mathcal{N}(A_T(x)u, I_V) \,\|\, \mathcal{N}(0, I_V)) = \frac{1}{2} u^\top M_T(x) u \tag{3}$$

The mutual information of this Gaussian channel is available in closed form (Eq. 4).

$$\boxed{\mathrm{GIF}_T(\tau; x) := \frac{1}{2} \log\det(I_{d_T} + \tau M_T(x))}. \tag{4}$$

GIF is therefore not merely a sensitivity or attribution score: it is the mutual information of an explicit channel constructed to match the local output geometry of the LLM.

2.3 Faithfulness and Local Soundness

Eq. 2 characterizes the infinitesimal geometry of the true model. We next show that the Gaussian surrogate remains faithful uniformly over perturbation directions within a fixed neighborhood.

Theorem 2.2 (Uniform second-order faithfulness)**.** *Under Assumption 2.1, let* $R := \min\{1, r_T(x)\}$. *There exists a computable constant* $C^\star$*, depending on* $\|W J_T(x)\|_{\mathrm{op}}$, $\|W\|_{\mathrm{op}}$, $r_T(x)$, $p(x)$, *and* $K_T(x)$*, such that for every* $\varepsilon \in \mathbb{R}$ *and every* $\|r\|_2 \leq R$*:*

$$\left| D_{\mathrm{KL}}\big(\pi_T(x_T + \varepsilon r) \| p(x)\big) - \frac{1}{2}\varepsilon^2 r^\top M_T(x) r \right| \leq C^\star |\varepsilon|^3.$$

The bound is uniform over perturbation directions: a single cubic remainder controls the approximation throughout the validity neighborhood. The guarantee is local and does not extend to arbitrary discrete prompt substitutions outside this neighborhood.

Most importantly, the operational GIF measure is conservative with respect to the true perturbation-to-output information flow.

Theorem 2.3 (Local soundness of GIF)**.** *Under Assumption 2.1:*

$$I(U_T; Y) \le \mathrm{GIF}_T(\tau; x) + o(\tau), \qquad \tau \to 0^+.$$

The proof proceeds in three steps. Replacing the intractable output marginal with the base distribution $p(x)$ first gives a conservative upper bound on the true mutual information. Theorem 2.2 then reduces the resulting expected KL to the Gaussian average of the local quadratic form, yielding $\frac{\tau}{2} \operatorname{tr} M_T(x) + o(\tau)$. Finally, this first-order trace term differs from the Gaussian-channel capacity in Eq. 4 only by $O(\tau^2)$.

2.4 Efficient Influence Proxy

Computing the log-determinant in Eq. 4 directly is expensive at LLM scale. Its leading-order term depends only on the trace of the Fisher pullback. We therefore define the *local influence* $\mathrm{Inf}_T(x) := \operatorname{tr} M_T(x)$, for which:

$$\mathrm{GIF}_T(\tau; x) = \frac{\tau}{2}\mathrm{Inf}_T(x) + O(\tau^2), \qquad \mathrm{GIF}_T(\tau; x) \le \frac{\tau}{2}\mathrm{Inf}_T(x). \tag{5}$$

Thus $\mathrm{Inf}_T(x)$ determines both the first-order growth rate of GIF and its global linear upper bound. We use it as a computationally more efficient proxy for GIF. § 3 shows how to estimate this trace efficiently using automatic differentiation and randomized low-rank trace estimation, without materializing $J_T(x)$ or $M_T(x)$.

3 Scalable Flow Estimation

§ 2 reduces scalable flow measurement to the local influence $\mathrm{Inf}_T(x) = \operatorname{tr} M_T(x)$, the first-order proxy for GIF. Computing this quantity naively is prohibitively expensive: explicitly forming $M_T(x)$ requires the span Jacobian $J_T(x)$, whose construction by reverse-mode automatic differentiation requires a backward pass for each hidden-state dimension (Baydin et al., 2018). We instead estimate the trace directly, without materializing either $J_T(x)$ or $M_T(x)$.

For a random probe z satisfying $\mathbb{E}[zz^\top] = I_{d_T}$, the Hutchinson identity (Hutchinson, 1989) expresses the local influence as an expectation over random quadratic forms (Eq. 6).

$$\mathrm{Inf}_T(x) = \operatorname{tr} M_T(x) = \mathbb{E}_z\left[z^\top M_T(x) z\right]. \tag{6}$$

We estimate this trace using Hutch++ (Meyer et al., 2021), which combines a randomized low-rank sketch with stochastic estimation of the residual. Crucially, the estimator requires only matrix–vector products with $M_T(x)$ rather than explicit access to the matrix itself.

Recall that $M_T(x) = J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x)$. Each required matrix–vector product therefore factorizes through the model Jacobian and output Fisher geometry (Eq. 7).

$$M_T(x) z = J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x) z. \tag{7}$$

This product can be evaluated without constructing the Jacobian. We first compute $J_T(x)z$ with a Jacobian–vector product (JVP), propagate the result through the language-model head and softmax Fisher operation, and finally multiply by $J_T(x)^\top$ with a vector–Jacobian product (VJP).

Since $M_T(x) \succeq 0$, Hutch++ provides a high-probability multiplicative approximation to $\mathrm{Inf}_T(x)$. Combining its upper confidence bound with Theorem 2.3 yields a corresponding probabilistic soundness guarantee. We give the exact estimator, query-complexity bound, and proof in Appendix C. § 4 evaluates how closely the resulting low-rank estimates preserve exact flow scores and source-token rankings.

4 Evaluation

We evaluate integrity violations on AgentDojo (Debenedetti et al., 2024) and MSB (Zhang et al., 2026a), which contain prompt-injection attacks against tool-using agents, and confidentiality violations on AgentDAM (Zharmagambetov et al., 2025), which evaluates privacy leakage in WebArena-derived browser tasks. We generate trajectories with six open-weight models spanning multiple

Table 1: Token-level localization performance swept over PSP@5–PSP@15 (shown as @5–@15). BM25 and Random rank non-overlapping 10-token spans; Random is averaged over 50 deterministic seeds. Per (dataset, model) block and metric column the best value is green and the worst red. Values are rounded to two decimals (leading zero dropped) to save space. Models: GPT-A/B = GPT-OSS-120B/20B; Gemma4 = Gemma-4-31B; Qwen-A/B/C = Qwen3-32B/14B/8B.

	Method	AgentDojo (Integrity)						MSB (Integrity)						AgentDAM (Confidentiality)					
		AUROC			AUPRC			AUROC			AUPRC			AUROC			AUPRC		
		@5	@10	@15	@5	@10	@15	@5	@10	@15	@5	@10	@15	@5	@10	@15	@5	@10	@15
GPT-A	GIF	**.79**	**.83**	**.85**	**.62**	**.69**	**.74**	.90	**.91**	**.93**	**.95**	**.95**	.96	**.69**	**.70**	**.74**	**.41**	**.40**	**.43**
	GIF$^-$	.77	.81	.83	.61	.66	.70	.89	**.91**	**.93**	.94	**.95**	.96	.66	.66	.65	.25	.22	.20
	RTBAS-LLA	.51	.51	.52	.24	.24	.25	.51	.53	.58	.75	.76	.79	.50	.50	.47	.18	.18	.17
	BM25	.61	.64	.70	.39	.42	.50	**.91**	**.91**	**.93**	**.95**	**.95**	**.97**	.50	.50	.51	.18	.18	.19
	Random	.51	.52	.53	.24	.25	.25	.57	.59	.62	.80	.81	.83	.53	.53	.54	.19	.19	.20
GPT-B	GIF	**.77**	**.85**	**.87**	**.45**	**.57**	**.61**	**.76**	**.80**	.82	**.81**	**.85**	**.85**	.67	.71	.72	.37	.43	.45
	GIF$^-$	.75	.83	**.87**	.44	.55	.60	.74	.78	**.83**	.79	.82	**.85**	.62	.69	.70	.31	.42	.46
	RTBAS-LLA	.74	.79	**.87**	.42	.45	.55	.62	.68	.73	.72	.76	.78	**.69**	**.75**	**.75**	**.43**	**.51**	**.51**
	BM25	.67	.70	.77	.35	.38	.48	.75	.76	.82	.80	.80	**.85**	.50	.50	.50	.12	.12	.12
	Random	.51	.52	.53	.15	.16	.16	.57	.58	.62	.69	.70	.73	.51	.52	.53	.13	.14	.14
Gemma4	GIF	.56	.62	.65	.19	.24	.26	.81	.83	.84	.81	.82	.83	**.59**	.58	.60	.13	.12	.14
	GIF$^-$	.55	.60	.62	.18	.22	.23	.80	.81	.83	.79	.80	.82	.58	**.62**	**.61**	.14	.14	.13
	RTBAS-LLA	.56	**.72**	**.74**	.22	.40	**.47**	.51	.56	.70	.59	.65	.76	.50	.50	.50	.15	.15	.15
	BM25	**.65**	.67	.71	**.38**	**.41**	**.47**	**.87**	**.88**	**.94**	**.89**	**.90**	**.95**	.50	.50	.50	.15	.15	.15
	Random	.52	.55	.57	.16	.18	.19	.62	.65	.69	.68	.70	.73	.50	.50	.51	**.16**	**.17**	**.17**
Qwen-A	GIF	**.81**	**.88**	**.92**	**.61**	**.71**	**.74**	.82	.84	.86	.92	.93	.94	**.61**	**.65**	**.69**	**.47**	**.49**	**.52**
	GIF$^-$	.79	.87	.91	.58	.69	**.74**	.82	.85	.87	.92	.93	.95	.58	.63	.67	.45	.48	.51
	RTBAS-LLA	.51	.56	.62	.18	.25	.30	.56	.57	.59	.78	.79	.80	.53	.54	.55	.45	.46	.47
	BM25	.68	.68	.70	.44	.45	.47	**.88**	**.88**	**.93**	**.94**	**.94**	**.97**	.50	.50	.49	.41	.41	.41
	Random	.52	.53	.55	.19	.20	.21	.55	.56	.59	.77	.78	.79	.54	.55	.57	.43	.44	.45
Qwen-B	GIF	**.78**	**.87**	**.90**	**.61**	**.75**	**.80**	.91	**.95**	**.97**	**.92**	**.96**	**.97**	**.63**	**.69**	.68	**.43**	**.43**	**.43**
	GIF$^-$	.76	.83	.87	.57	.68	.74	.90	.94	**.97**	.90	.94	**.97**	.61	.66	**.70**	.38	.41	**.43**
	RTBAS-LLA	.51	.60	.72	.19	.33	.52	.49	.49	.51	.53	.53	.54	.50	.50	.50	.31	.31	.31
	BM25	.72	.72	.76	.54	.54	.60	**.92**	.92	.95	**.92**	.92	.95	.50	.50	.50	.31	.31	.31
	Random	.53	.54	.56	.20	.21	.23	.56	.57	.60	.59	.60	.63	.51	.51	.51	.32	.32	.33
Qwen-C	GIF	**.76**	**.85**	.87	.44	**.57**	.61	.85	.89	.92	.88	.92	.94	**.65**	**.69**	**.71**	.39	.41	.42
	GIF$^-$	**.76**	.83	**.89**	**.46**	.55	**.64**	**.89**	**.92**	**.94**	**.91**	**.93**	**.95**	.61	.68	**.71**	**.40**	**.44**	**.44**
	RTBAS-LLA	.51	.59	.67	.12	.23	.35	.49	.52	.57	.59	.60	.63	.50	.50	.50	.28	.28	.28
	BM25	.67	.67	.70	.37	.36	.40	.86	.87	**.94**	.89	.90	**.95**	.50	.51	.50	.28	.29	.29
	Random	.52	.53	.55	.12	.13	.14	.53	.54	.57	.62	.62	.64	.51	.51	.51	.29	.29	.30
Average	GIF	**.75**	**.82**	**.85**	**.49**	**.59**	**.63**	.84	**.87**	.89	.88	**.90**	.91	**.64**	**.67**	**.69**	**.37**	**.38**	**.40**
	GIF$^-$	.73	.79	.83	.47	.56	.61	.84	**.87**	.90	.88	**.90**	.92	.61	.66	.67	.32	.35	.36
	RTBAS-LLA	.56	.63	.69	.23	.32	.41	.53	.56	.61	.66	.68	.72	.54	.55	.55	.30	.32	.32
	BM25	.67	.68	.72	.41	.43	.49	**.86**	**.87**	**.92**	**.90**	**.90**	**.94**	.50	.50	.50	.24	.24	.24
	Random	.52	.53	.55	.18	.19	.20	.57	.58	.61	.69	.70	.73	.52	.52	.53	.26	.26	.27

sizes and families: Qwen 3 8B, Qwen 3 14B, Qwen 3 32B, Gemma 4 31B, GPT OSS 20B, and GPT OSS 120B. We use each benchmark's native trajectory-level oracle as ground truth. Full trajectory-generation, annotation, and compute details are given in Appendix E.

GIF assigns each source token in the preceding context an influence score relative to a generated sink token. For actions containing multiple sink tokens, we max-pool each source token's influence across the action to obtain a single source ranking used throughout the evaluation. Full labeling and aggregation details are given in Appendix D.

4.1 How well does GIF surface policy-relevant information flow?

We first evaluate GIF as a policy signal in isolation. Existing agent benchmarks provide trajectory-level violation labels and annotations of prompt-injection or private-information spans, but not ground-truth causal dependencies for individual generated actions. We therefore evaluate whether each attribution method ranks these policy-relevant source tokens more highly when the agent violates policy.

For each candidate action, we compute *Policy Source Precision at* k (PSP@k), the fraction of the top-k attributed source tokens that overlap benchmark-annotated policy-relevant content. For trajectories with multiple candidate actions, we take the maximum PSP@k as the trajectory-level score. We then sweep a decision threshold over this score and report AUROC and AUPRC. This evaluates the separation induced by the attribution ranking without assuming that raw scores are calibrated across methods. Annotation alignment and metric details are given in Appendix E.

Baselines. We compare GIF against baselines that isolate different sources of attribution signal. GIF$^-$ uses the same model gradients but measures sensitivity of the generated token's logit rather than the geometry of the full next-token distribution. RTBAS-LLA uses final-layer attention averaged across heads as a scalable attention-based baseline. Full RTBAS requires backpropagation through attention maps across layers and heads and is infeasible at our full model and context scales; we evaluate our reimplementation on a reduced setting in Appendix F. We additionally include BM25 and random-span controls to measure lexical overlap and chance selection, respectively.

GIF provides the strongest overall model-based localization signal. Averaged across models, GIF outperforms RTBAS-LLA on each benchmark and generally improves over GIF$^-$. The gains over GIF$^-$ are largest on AgentDojo and AgentDAM, indicating that modeling movement of the full next-token distribution provides useful signal beyond sensitivity of the realized-token logit alone. The gap to RTBAS-LLA is larger, showing that the final-layer attention signal is substantially less informative for policy-relevant source localization. BM25 is competitive on MSB, where many attacks copy paths, tool arguments, or other strings verbatim or nearly verbatim into downstream actions, making lexical overlap an unusually strong signal.

GIF remains competitive with frontier LLM attribution at much lower cost. We additionally compare against GPT-5.5 xhigh as an inference-time token-attribution judge on stratified samples from each benchmark. Averaged over the evaluated cutoffs, GIF achieves 11.2% higher AUROC and 12.0% higher AUPRC while reducing average attribution cost from \$147.90 to \$15.98, a $9.3\times$ reduction. Gemma 4 31B is the main exception, where GPT-5.5 obtains higher AUPRC. Full per-model results and sampling details are reported in Appendix F.

4.2 Can GIF support efficient LLM-based policy decisions?

Attribution alone does not determine whether an observed dependency violates policy: untrusted information may be legitimate task input, while private information may be necessary to satisfy an authorized request. We therefore test whether GIF can surface compact evidence for an LLM-based policy decision procedure. For each candidate action, GIF selects the $k \in \{1, 2, 3\}$ highest-influence source spans and provides them, together with the user task and target action, to an LLM that returns Allow or Deny. We evaluate Qwen 3 8B, GPT OSS 120B, Gemma 4 31B, and GPT-5.5 xhigh in this role. As a baseline, the same model judges the full trajectory under the same policy ruleset. We report detection quality and total token usage, including input, output, and reasoning tokens.

GIF-guided evidence preserves or improves decision quality. As shown on the left of Fig. 2, across models and metrics, GIF-k achieves $0.97\times$–$1.22\times$ the performance of the corresponding full-trajectory judge. The largest gains occur for Qwen 3 8B, whose F1 improves by 9–12 percentage points, suggesting that explicitly surfacing high-influence evidence is especially useful for smaller LLMs. On MSB, Qwen 3 8B with GIF comes within 1.04 percentage points of GPT-5.5 xhigh's F2; with GIF-3, it matches GPT-5.5 xhigh across AgentDojo and MSB while incurring more than $81\times$ lower model inference cost.

GIF substantially reduces the context required for policy decisions. As shown on the right of Fig. 2, on AgentDojo and MSB, GIF-1 uses only 0.27–$0.37\times$ as many tokens as the corresponding full-trajectory judge, reducing token usage by 63–73%. Even GIF-3 reduces token usage by 59–67%. Most of these savings come from input tokens, which account for 84–91% of the full-trajectory judge's token usage. The savings are smaller on AgentDAM, where GIF-k uses 0.75–$0.82\times$ as many tokens as the full-trajectory judge. Unlike the longer tool-use traces in AgentDojo and MSB, the WebArena-derived agent maintains much of its state in the browser environment rather than in conversational history. Consequently, there is less irrelevant trajectory context for GIF to remove.

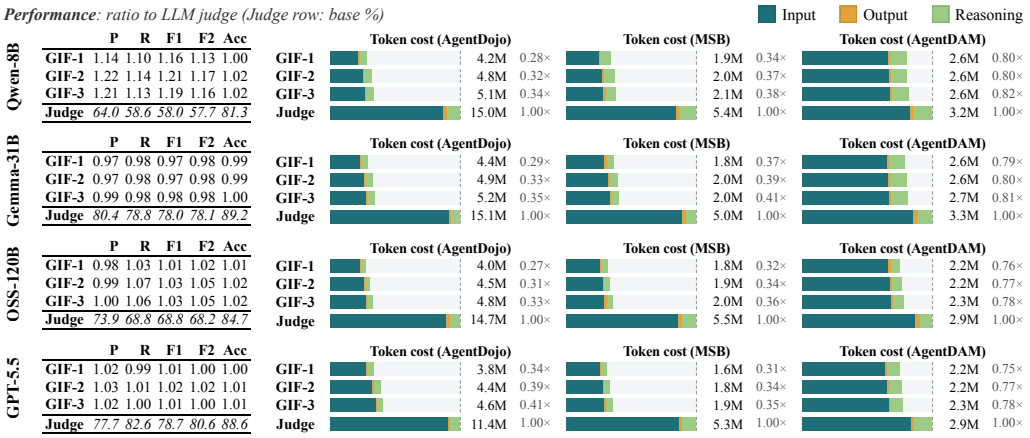


Figure 2: **GIF-guided policy decisions vs. full-trajectory judging.** Open-weight rows use the same model for trajectory generation and downstream policy decisions; the GPT-5.5 row pools trajectories across all six open-weight agent models and uses GPT-5.5 only for policy decisions. *Left:* detection quality of GIF-k, normalized by the same model's full-trajectory judge (judge row: absolute %); values ≥ 1 indicate that GIF-selected evidence matches or exceeds full-trajectory judging. *Right:* token usage per benchmark, split into input, output, and reasoning tokens; the gray factor is relative to the full-trajectory judge.

Table 2: Hutch++ sketched influence vs. exact influence as a function of sketch rank r, on 150 AgentDojo sink tokens (Qwen 3 0.6B analysis model). Score fidelity: MSE, relative L_2, normalized relative L_2. Ranking quality: Spearman, and top-k index overlap with the exact ranking.

Rank	MSE	Rel. L_2	Norm. rel. L_2	Spearman	Top-20	Top-50
1	30.531	0.634	0.335	0.957	0.826	0.855
2	15.214	0.545	0.279	0.977	0.860	0.889
4	5.886	0.467	0.250	0.983	0.878	0.904
8	0.174	0.413	0.240	0.986	0.890	0.915
16	0.077	0.384	0.229	0.987	0.897	0.920
32	0.059	0.370	0.224	0.988	0.898	0.922
64	0.056	0.359	0.218	0.988	0.898	0.924

4.3 How robust is GIF to practical approximations and variations?

We next examine the main approximations and design choices involved in using GIF in practice: low-rank flow estimation, use of surrogate analysis models, adversarial obfuscation of policy-relevant inputs, and the attribution cutoff used in the localization evaluation.

Low-rank estimation preserves source rankings at small rank. We compare approximate influence scores against exact Fisher-pullback traces on 150 AgentDojo sink tokens drawn from 30 Qwen 3 8B tool calls, with five generated tool-call tokens sampled from each call. Each sink token is analyzed with Qwen 3 0.6B, allowing us to compute the exact source-token influence vector and compare it with approximations at ranks $r \in \{1, 2, 4, 8, 16, 32, 64\}$ using five fixed seeds per rank.

As shown in Tab. 2, increasing rank substantially improves absolute score fidelity, while source-token rankings are already stable at very low rank. Mean MSE falls from 30.53 at r=1 to 0.17 at r=8 and 0.056 at r=64, whereas Spearman correlation is already 0.957 at r=1, rising only to 0.986 and 0.988 at r=8 and r=64. Top-50 overlap similarly increases from 0.855 to 0.915 and 0.924. Thus higher rank mainly improves absolute score fidelity; the ranking needed to surface high-influence evidence is preserved even by small sketches.

Small surrogate models preserve downstream policy performance. Computing GIF requires gradients through an analysis model, but that model need not be the model that generated the tra-

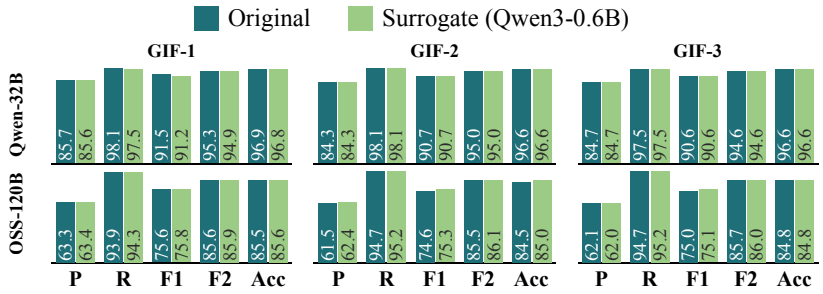


Figure 3: **Surrogate analysis models.** Detection precision (P), recall (R), F_1, F_2, and accuracy (Acc) for the three GIF-k evidence budgets when AgentDojo trajectories from Qwen 3 32B (top) and GPT OSS 120B (bottom) are attributed using the original target model (*Original*) versus a fixed Qwen 3 0.6B surrogate. The surrogate tracks the original across both targets and all variants, including the cross-family GPT OSS 120B target.

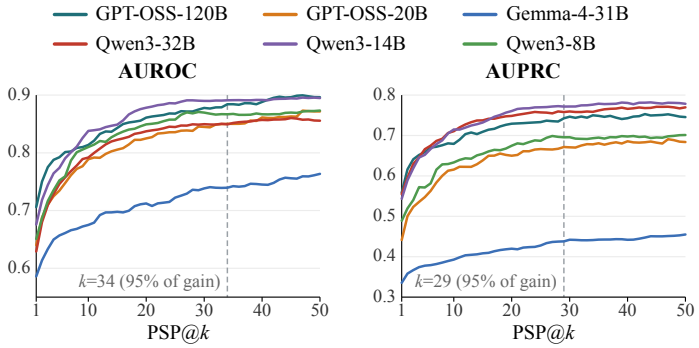


Figure 4: **Attribution cutoff k.** AUROC (left) and AUPRC (right) of GIF attribution as a function of the cutoff k in PSP@k, for six models averaged over AgentDojo, MSB, and AgentDAM. Both metrics saturate: 95% of the AUROC gain is reached by k=34 and of the AUPRC gain by k=29.

jectory. We reattribute AgentDojo trajectories generated by Qwen 3 32B and GPT OSS 120B using either the original trajectory model or a fixed Qwen 3 0.6B surrogate, while holding the downstream LLM-based policy decision procedure fixed (Fig. 3).

Across both target models and all three GIF variants, replacing the original analysis model with Qwen 3 0.6B changes F_1 by at most 0.7 percentage points and any reported metric by at most 0.9 points. The result also holds for the cross-family GPT OSS 120B target, which is approximately 200× larger than the surrogate. This suggests that GIF can be computed using inexpensive surrogate models even when gradients of the target model are unavailable or too expensive.

GIF remains informative under discrete prompt obfuscation. Theorem 2.2 is uniform over perturbation directions within its local validity neighborhood, but does not cover arbitrary discrete prompt transformations. We therefore additionally test GIF under base64 and hexadecimal encoding, character spacing, chunking, reversal, and paraphrasing of prompt injections in AgentDojo. At PSP@50, GIF achieves average AUROC/AUPRC of 0.649/0.471 across six models, compared with 0.473/0.371 for BM25. The gap under obfuscation suggests that GIF captures model-mediated influence beyond surface lexical overlap.

Localization performance is not sensitive to a narrow choice of attribution cutoff. Finally, we vary the cutoff k in PSP@k from 1 to 50 while holding the benchmark-defined policy-token set fixed. Averaged across the six models and three benchmarks, AUROC rises from 0.66 at k=1 to 0.87 at k=50, while AUPRC rises from 0.50 to 0.70 (Fig. 4). Most improvement occurs well before the end of the sweep: 95% of the total AUROC gain is reached by k=34 and of the AUPRC gain by k=29. The main localization results are therefore not tied to an unusually narrow cutoff choice.

4.4 What flow patterns does GIF reveal in representative agent trajectories?

GIF reveals policy-relevant dependencies that trajectory-level success labels can miss. In an AgentDojo email task, an injected instruction causes the agent to send to the attacker's address but not the requested body; although the benchmark marks the attack unsuccessful, GIF attributes 31% of the measured flow into the recipient to the injected prompt. Conversely, in an AgentDAM confidentiality example, GIF assigns zero measured local influence to a private address that is absent from the public output, while RTBAS-LLA assigns it 9%. These examples illustrate how GIF can surface partial compliance while avoiding over-tainting irrelevant context. Full trajectories and visualizations are provided in Appendix G.

5 Related Work

Traditional IFC. Information-flow control tracks confidential or untrusted data through programs and systems using security labels, noninterference-style policies, and explicit declassification (Denning, 1976; Sabelfeld & Myers, 2003; Volpano et al., 1996; Myers, 1999; Sabelfeld & Sands, 2009). Noninterference and related information-flow policies are naturally relational, or hyperproperties (Clarkson & Schneider, 2010); quantitative information flow (QIF) further measures how much information is leaked (Smith, 2009; Backes et al., 2009; Chatzikokolakis et al., 2010; Chothia & Guha, 2011). Static and abstract-interpretation techniques can conservatively compute or approximate such flows (Clark et al., 2007; Cousot & Cousot, 1977). These foundations, however, do not specify how information propagates through an LLM invocation, where heterogeneous inputs mix inside continuous neural computation.

Agentic IFC. Recent systems adapt IFC and related security abstractions to LLM agents through planning/execution separation, information-flow enforcement, privilege control, and tool-call screening (Wu et al., 2024; Costa et al., 2025; Debenedetti et al., 2025; Beurer-Kellner et al., 2025; Shi et al., 2026; Zhong et al., 2025; Siddiqui et al., 2025). These approaches largely control flow around the model rather than providing a quantitative semantics for flow through its internal computation. GIF is complementary: it supplies fine-grained model-mediated dependencies for downstream enforcement or declassification.

Heuristic Interpretability and Attribution. Prior work estimates input influence using gradients, saliency, integrated gradients, attention, influence functions, and related attribution methods (Simonyan et al., 2014; Sundararajan et al., 2017; Jain & Wallace, 2019; Koh & Liang, 2017; Wallace et al., 2020). Recent LLM work extends attribution to token- and data-level influence, including Jacobian- and gradient-based approaches (Liu et al., 2026; Chen et al., 2026; Jiao et al., 2025). GIF instead gives local Jacobian geometry a quantitative information-flow interpretation through a Gaussian surrogate channel, together with a local soundness theorem relating its computable score to true perturbation-to-output flow.

Prompt-Injection and Privacy-Leakage Defenses. Prompt-injection defenses include input transformations, privilege control, adaptive-attack-aware defenses, intent analysis, and context purification (Hines et al., 2024; Shi et al., 2025; 2026; Zhan et al., 2025; Kang et al., 2025; Zhang et al., 2026b). Privacy-focused work studies leakage of personal information and privacy failures in web and multi-agent systems (Alizadeh et al., 2025; Zharmagambetov et al., 2025; El Yagoubi et al., 2026). These approaches are largely task-specific defenses or evaluations; GIF instead provides a general quantitative semantics for tracing model-mediated influence from private or untrusted spans to downstream outputs and actions.

6 Conclusion

We introduced Geometric Information Flow (GIF), a quantitative semantics for information flow through LLMs. GIF models influence as a local channel from input spans to downstream actions. Our results support GIF as a foundation for fine-grained IFC in agentic systems, with strong policy-detection performance and substantially reduced downstream inference cost.

AI USE STATEMENT

We used generative AI tools to assist with polishing the manuscript's prose and grammar, as a coding assistant during software development, and as a Lean 4 proof-writing assistant. All AI-assisted code and proof scripts were reviewed by the authors, and the Lean 4 formalization was mechanically checked by the Lean proof assistant. We take responsibility for the final content of this work, including all text, code, and formal proofs produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the code and Lean 4 formalization for GIF on our anonymous project website `https://geomifc.github.io`. Our evaluation uses publicly available benchmarks (AgentDojo, MSB, and AgentDAM) and open-weight models, with compute environment documented in the appendix. We additionally report full per-model results, approximation studies, and reduced-scale comparisons for baselines that cannot be run at full scale under our compute budget. Together, these materials are intended to enable reproduction of both the theoretical results and the empirical evaluation.

REFERENCES

Meysam Alizadeh, Zeynab Samei, Daria Stetsenko, and Fabrizio Gilardi. Simple prompt injection attacks can leak personal data observed by llm agents during task execution, 2025. URL `https://arxiv.org/abs/2506.01055`.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Michael Backes, Boris Köpf, and Andrey Rybalchenko. Automatic discovery and quantification of information leaks. In *2009 30th IEEE Symposium on Security and Privacy*, pp. 141–153. IEEE, 2009.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.

Luca Beurer-Kellner, Ana-Maria Creţu, Beat Buesser, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew Paverd, Florian Tramèr, and Václav Volhejn. Design Patterns for Securing LLM Agents against Prompt Injections, June 2025. URL `http://arxiv.org/abs/2506.08837`. arXiv:2506.08837 version: 1.

Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical measurement of information leakage. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pp. 390–404. Springer, 2010.

Jianhui Chen, Yuzhang Luo, and Liangming Pan. Mechanistic data attribution: Tracing the training origins of interpretable llm units, 2026. URL `https://arxiv.org/abs/2601.21996`.

Tom Chothia and Apratim Guha. A statistical test for information leaks using continuous mutual information. In *2011 IEEE 24th Computer Security Foundations Symposium*, pp. 177–190. IEEE, 2011.

David Clark, Sebastian Hunt, and Pasquale Malacaria. A static analysis for quantifying information flow in a simple imperative language. *Journal of Computer Security*, 15(3):321–371, 2007.

Michael R. Clarkson and Fred B. Schneider. Hyperproperties. *J. Comput. Secur.*, 18(6):1157–1210, September 2010. ISSN 0926-227X.

Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing ai agents with information-flow control. *CoRR*, abs/2505.23643, 2025.

Patrick Cousot and Radhia Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of the 4th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, POPL '77, pp. 238–252, New York, NY, USA, 1977. Association for Computing Machinery. ISBN 9781450373500. doi: 10.1145/512950.512973. URL https://doi.org/10.1145/512950.512973.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=m1YYAQjO3w.

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, June 2025. URL http://arxiv.org/abs/2503.18813. arXiv:2503.18813.

Dorothy E Denning. A lattice model of secure information flow. *Communications of the ACM*, 19 (5):236–243, 1976.

Faouzi El Yagoubi, Godwin Badu-Marfo, and Ranwa Al Mallah. Agentleak: A full-stack benchmark for privacy leakage in multi-agent llm systems. *arXiv preprint arXiv:2602.11510*, 2026. URL https://arxiv.org/abs/2602.11510. Submitted to arXiv on 12 Feb 2026.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting, 2024. URL https://arxiv.org/abs/2403.14720.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Cathy Jiao, Yijun Pan, Emily Xiao, Daisy Sheng, Niket Jain, Hanzhang Zhao, Ishita Dasgupta, Jiaqi W. Ma, and Chenyan Xiong. Date-lm: Benchmarking data attribution evaluation for large language models, 2025. URL https://arxiv.org/abs/2507.09424.

Mintong Kang, Chong Xiang, Sanjay Kariyappa, Chaowei Xiao, Bo Li, and Edward Suh. Mitigating indirect prompt injection via instruction-following intent analysis, 2025. URL https://arxiv.org/abs/2512.00966.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Toni J. B. Liu, Baran Zadeoğlu, Nicolas Boullé, Raphaël Sarfati, and Christopher J. Earls. Jacobian scopes: token-level causal attributions in llms, 2026. URL https://arxiv.org/abs/2601.16407.

Raphael A Meyer, Cameron Musco, Christopher Musco, and David P Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 142–155. SIAM, 2021.

Andrew C. Myers. Jflow: practical mostly-static information flow control. In *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '99, pp. 228–241, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130953. doi: 10.1145/292540.292561. URL https://doi.org/10.1145/292540.292561.

Andrei Sabelfeld and Andrew C Myers. Language-based information-flow security. *IEEE Journal on selected areas in communications*, 21(1):5–19, 2003.

Andrei Sabelfeld and David Sands. Declassification: Dimensions and principles. *Journal of Computer Security*, 17(5):517–548, October 2009. ISSN 18758924, 0926227X. doi: 10.3233/JCS-2009-0352. URL https://journals.sagepub.com/doi/full/10.3233/JCS-2009-0352.

Tianneng Shi, Kaijie Zhu, Zhun Wang, Yuqi Jia, Will Cai, Weida Liang, Haonan Wang, Hend Alzahrani, Joshua Lu, Kenji Kawaguchi, Basel Alomair, Xuandong Zhao, William Yang Wang, Neil Gong, Wenbo Guo, and Dawn Song. Promptarmor: Simple yet effective prompt injection defenses, 2025. URL https://arxiv.org/abs/2507.15219.

Tianneng Shi, Jingxuan He, Zhun Wang, Hongwei Li, Linyu Wu, Wenbo Guo, and Dawn Song. Progent: Securing ai agents with privilege control, 2026. URL https://arxiv.org/abs/2504.11703.

Shoaib Ahmed Siddiqui, Radhika Gaonkar, Boris Köpf, David Krueger, Andrew Paverd, Ahmed Salem, Shruti Tople, Lukas Wutschitz, Menglin Xia, and Santiago Zanella-Beguelin. Permissive information-flow analysis for large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ufYRO8y3mr.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Geoffrey Smith. On the foundations of quantitative information flow. In *International Conference on Foundations of Software Science and Computational Structures*, pp. 288–302. Springer, 2009.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328. JMLR.org, 2017.

Dennis Volpano, Cynthia Irvine, and Geoffrey Smith. A sound type system for secure flow analysis. *Journal of computer security*, 4(2-3):167–187, 1996.

Eric Wallace, Matt Gardner, and Sameer Singh. Interpreting predictions of NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 20–23, 2020.

Fangzhou Wu, Ethan Cecchetti, and Chaowei Xiao. System-level defense against indirect prompt injection attacks: An information flow control perspective. *CoRR*, abs/2409.19091, 2024.

Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break defenses against indirect prompt injection attacks on llm agents, 2025. URL https://arxiv.org/abs/2503.00061.

Dongsen Zhang, Zekun Li, Xu Luo, Xuannan Liu, Pei Pei Li, and Wenjun Xu. MCP security bench (MSB): Benchmarking attacks against model context protocol in LLM agents. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=irxxkFMrry.

Tian Zhang, Yiwei Xu, Juan Wang, Keyan Guo, Xiaoyang Xu, Bowen Xiao, Quanlong Guan, Jinlin Fan, Jiawei Liu, Zhiquan Liu, and Hongxin Hu. Agentsentry: Mitigating indirect prompt injection in llm agents via temporal causal diagnostics and context purification, 2026b. URL https://arxiv.org/abs/2602.22724.

Arman Zharmagambetov, Chuan Guo, Ivan Evtimov, Maya Pavlova, Ruslan Salakhutdinov, and Kamalika Chaudhuri. AgentDAM: Privacy leakage evaluation for autonomous web agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qaxf7q41aK.

Peter Yong Zhong, Siyuan Chen, Ruiqi Wang, McKenna McCall, Ben L. Titzer, Heather Miller, and Phillip B. Gibbons. Rtbas: Defending llm agents against prompt injection and privacy leakage, 2025. URL https://arxiv.org/abs/2502.08966.

A Information-Theoretic Preliminaries

We collect the information-theoretic facts used in our analysis. Throughout, $\mathcal{L}_X$ denotes the probability law of a random variable X, and $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance Σ.

A.1 Kullback–Leibler Divergence

Kullback–Leibler (KL) divergence measures how distinguishable one distribution is from another. For probability distributions P and Q on a common measurable space,

$$D_{\mathrm{KL}}(P\|Q) := \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right],$$

with the convention that $D_{\mathrm{KL}}(P\|Q) = +\infty$ when the likelihood ratio is not well-defined under P.

For discrete distributions with probability vectors p and q,

$$D_{\mathrm{KL}}(P\|Q) = \sum_i p_i \log \frac{p_i}{q_i},$$

with the usual conventions $0 \log(0/q_i) = 0$ and $p_i \log(p_i/0) = +\infty$ for $p_i > 0$. KL divergence is nonnegative and vanishes exactly when the two distributions are equal.

A.2 Fisher Information

For a parametric distribution P_η with parameter η, Fisher information measures how sensitively the distribution changes under infinitesimal perturbations of η. If $Y \sim P_\eta$, the *score*

$$\nabla_\eta \log P_\eta(Y)$$

measures how the log-probability of the sampled outcome changes with the parameter. The Fisher information matrix is

$$F(\eta) := \mathbb{E}_{Y\sim P_\eta}\left[\nabla_\eta \log P_\eta(Y) \nabla_\eta \log P_\eta(Y)^\top\right].$$

Fisher information gives the local second-order geometry of KL divergence (Amari & Nagaoka, 2000). Under standard smoothness and fixed-support regularity conditions,

$$D_{\mathrm{KL}}(P_\eta \| P_{\eta+\delta}) = \frac{1}{2}\delta^\top F(\eta)\delta + o(\|\delta\|^2). \tag{8}$$

In this work, Fisher information appears through the softmax family. Let $z \in \mathbb{R}^V$ be a vector of logits and $p = \mathrm{softmax}(z)$. If the sampled token is $Y = i$, then the score with respect to the logits is $e_i - p$, where e_i is the ith standard basis vector. Hence

$$\begin{aligned} F_{\mathrm{sm}}(p) &= \mathbb{E}_{Y\sim p}\left[(e_Y - p)(e_Y - p)^\top\right] \\ &= \sum_{i=1}^{V} p_i (e_i - p)(e_i - p)^\top \\ &= \mathrm{diag}(p) - pp^\top. \end{aligned} \tag{9}$$

A.3 Mutual Information

Mutual information measures how much one random variable reveals about another. For jointly distributed random variables (U, Y),

$$I(U;Y) := D_{\mathrm{KL}}\big(\mathcal{L}_{(U,Y)} \,\big\|\, \mathcal{L}_U \otimes \mathcal{L}_Y\big)\,. \tag{10}$$

The product law $\mathcal{L}_U \otimes \mathcal{L}_Y$ represents the hypothetical distribution in which U and Y have their original marginals but are independent. Mutual information is therefore nonnegative and vanishes exactly when U and Y are independent.

A.4 Linear Gaussian Channels

A linear Gaussian channel maps a Gaussian input through a linear transformation and adds independent Gaussian noise. Let

$$U \sim \mathcal{N}(0, \tau I_r), \qquad N \sim \mathcal{N}(0, I_s),$$

independently, and let

$$Y = AU + N$$

for $A \in \mathbb{R}^{s\times r}$.

The mutual information of this channel has the closed form

$$\begin{aligned} I(U;Y) &= \frac{1}{2}\log\det\big(I_s + \tau A A^\top\big) \\ &= \frac{1}{2}\log\det\big(I_r + \tau A^\top A\big)\,, \end{aligned} \tag{11}$$

where the second equality follows from Sylvester's determinant identity. Thus the information transmitted by the channel is controlled by the spectrum of $A^\top A$.

B Geometric Information Flow: Full Theory and Proofs

This appendix gives the complete mathematical development underlying § 2. All mathematical claims are formalized and checked in Lean 4; the formal development is available on our anonymous website: `https://geomifc.github.io`.

B.1 Detailed Setup and True Information Flow

B.1.1 Prompt, Span, and Next-Token Map

Let V be the vocabulary size, n the prompt length, d the token-embedding dimension, and m the hidden-state dimension at a selected output position $t^\star$.

We represent the fixed prompt by its sequence of token embeddings $x = (x_1, \dots, x_n) \in (\mathbb{R}^d)^n$. Let $T \subseteq \{1, \dots, n\}$ be a token span, write $d_T := |T|d$, and let $x_T \in \mathbb{R}^{d_T}$ denote the concatenation of the embeddings in the span.

For $\xi \in \mathbb{R}^{d_T}$, $\mathrm{perturb}_T(x, \xi)$ denotes the prompt obtained by replacing x_T by ξ while holding every off-span coordinate fixed. In particular, $\mathrm{perturb}_T(x, x_T) = x$.

Let $h_\theta : (\mathbb{R}^d)^n \to \mathbb{R}^m$ be the fixed-parameter map from prompt embeddings to the hidden state at position $t^\star$. Restricting the model to perturbations of the span gives

$$h_T(\xi) := h_\theta(\mathrm{perturb}_T(x, \xi))\,,$$

so that

$$h_T(x_T) = h_\theta(x).$$

Let $W \in \mathbb{R}^{V\times m}$ and $b \in \mathbb{R}^V$ be the language-model head. The corresponding next-token distribution is

$$\pi_T(\xi) := \mathrm{softmax}(W h_T(\xi) + b) \in \Delta^{V-1},$$

where

$$\Delta^{V-1} := \left\{ q \in \mathbb{R}^V : q_i \geq 0,\ \sum_i q_i = 1 \right\}.$$

We write

$$p(x) := \pi_T(x_T)$$

for the next-token distribution at the original prompt.

The Jacobian of the span-restricted hidden-state map is

$$J_T(x) := Dh_T(x_T) \in \mathbb{R}^{m \times d_T}.$$

For a sufficiently small perturbation u, the product $J_T(x)u$ is the first-order displacement of the hidden state induced by perturbing the span along u.

B.1.2 Local Regularity

Our analysis is local around the observed prompt. We use the local regularity condition stated in Assumption 2.1: there exist constants $r_T(x) > 0$ and $K_T(x) \geq 0$ such that, for every $\delta \in \mathbb{R}^{d_T}$ with $\|\delta\|_2 \leq r_T(x)$,

$$\|h_T(x_T + \delta) - h_T(x_T) - J_T(x)\delta\|_2 \leq K_T(x)\|\delta\|_2^2. \tag{12}$$

Thus the Jacobian captures the first-order hidden-state displacement, while the residual is quadratic in the perturbation magnitude.

B.1.3 True Perturbation-to-Output Information Flow

Fix a perturbation scale $\tau > 0$ and let

$$U_T \sim \mathcal{N}(0, \tau I_{d_T})$$

be an isotropic perturbation of the source span. Conditional on $U_T = u$, the model samples the next token Y according to

$$\Pr(Y = i \mid U_T = u) = \pi_T(x_T + u)_i, \qquad i = 1, \ldots, V.$$

Definition B.1 (True local information flow)**.** The true perturbation-to-output information flow at scale τ is

$$I(U_T; Y) := D_{\mathrm{KL}}\big(\mathcal{L}_{(U_T,Y)} \,\big\|\, \mathcal{L}_{U_T} \otimes \mathcal{L}_Y\big). \tag{13}$$

The quantity is zero exactly when the sampled output token is independent of the span perturbation, and increases as the perturbation becomes more observable through the model output.

B.1.4 Exact Form of the True Information Flow

The mutual information in Definition B.1 admits an exact characterization in terms of the perturbed output distributions.

Theorem B.2 (Exact form of the true information flow)**.** *Let*

$$P_Y := \mathbb{E}_{U_T}\left[\pi_T(x_T + U_T)\right] \in \Delta^{V-1}$$

denote the output marginal. Then

$$I(U_T; Y) = \mathbb{E}_{U_T}\left[D_{\mathrm{KL}}(\pi_T(x_T + U_T) \,\|\, P_Y)\right]. \tag{14}$$

Equivalently,

$$I(U_T; Y) = H(P_Y) - \mathbb{E}_{U_T}\left[H(\pi_T(x_T + U_T))\right], \tag{15}$$

where

$$H(q) := -\sum_y q_y \log q_y.$$

Proof. For a fixed perturbation u, let

$$\pi_u := \pi_T(x_T + u).$$

By definition of the conditional law,

$$\Pr(Y = y \mid U_T = u) = \pi_u(y),$$

while

$$P_Y(y) = \mathbb{E}_{U_T}[\pi_{U_T}(y)].$$

Expanding mutual information using the conditional-to-marginal likelihood ratio,

$$\begin{aligned} I(U_T; Y) &= \mathbb{E}_{U_T}\left[\sum_y \pi_{U_T}(y) \log \frac{\pi_{U_T}(y)}{P_Y(y)}\right] \\ &= \mathbb{E}_{U_T}\left[D_{\mathrm{KL}}(\pi_T(x_T + U_T) \| P_Y)\right], \end{aligned}$$

which proves Eq. 14.

For the entropy form,

$$D_{\mathrm{KL}}(\pi_{U_T} \| P_Y) = -H(\pi_{U_T}) - \sum_y \pi_{U_T}(y) \log P_Y(y).$$

Taking expectation over U_T and using

$$\mathbb{E}_{U_T}[\pi_{U_T}(y)] = P_Y(y),$$

the second term becomes

$$-\sum_y P_Y(y) \log P_Y(y) = H(P_Y),$$

yielding Eq. 15. □

Although Theorem B.2 characterizes the desired quantity exactly, it is not operational: evaluating P_Y requires integrating the nonlinear model output over the perturbation distribution. We therefore construct a tractable local surrogate.

B.2 Local KL Geometry

B.2.1 Output Geometry

Because the LLM output is a probability distribution, the appropriate local geometry is given by the softmax Fisher matrix $F_{\mathrm{sm}}(p(x))$ from Eq. 9. For a small logit perturbation $\delta z \in \mathbb{R}^V$,

$$\delta z^\top F_{\mathrm{sm}}(p(x)) \delta z$$

is the local squared Fisher–Rao displacement of the next-token distribution. Equivalently, to second order it is twice the KL divergence between the base and perturbed outputs.

B.2.2 Fisher Pullback to the Input Span

For a small span perturbation $u \in \mathbb{R}^{d_T}$, the first-order hidden state displacement is

$$J_T(x) u,$$

and the corresponding first-order logit displacement is

$$W J_T(x) u.$$

Substituting this displacement into the output Fisher geometry gives

$$\begin{aligned} &(W J_T(x) u)^\top F_{\mathrm{sm}}(p(x)) (W J_T(x) u) \\ &= u^\top M_T(x) u, \end{aligned}$$

where

$$M_T(x) := J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x) \in \mathbb{R}^{d_T \times d_T}. \tag{16}$$

We call $M_T(x)$ the *Fisher pullback*: it transports the local output geometry back onto the source-span coordinates. Thus $u^\top M_T(x) u$ is large when moving the span along u induces a large local movement of the next-token distribution.

B.2.3 Local KL Curvature

Lemma B.3 (Local KL geometry)**.** *Under Assumption 2.1, as $u \to 0$,*

$$D_{\mathrm{KL}}(\pi_T(x_T+u) \,\|\, p(x)) = \frac{1}{2} u^\top M_T(x) u + o(\|u\|_2^2). \tag{17}$$

Equivalently, $M_T(x)$ is the Hessian at $u = 0$ of

$$u \mapsto D_{\mathrm{KL}}(\pi_T(x_T+u) \| p(x)) \,.$$

Proof. Let

$$z_0 := W h_T(x_T) + b$$

be the base logits, and define the induced logit perturbation

$$\Delta(u) := W\big(h_T(x_T+u) - h_T(x_T)\big).$$

By differentiability of h_T at x_T,

$$h_T(x_T+u) - h_T(x_T) = J_T(x)u + o(\|u\|),$$

and hence

$$\Delta(u) = W J_T(x) u + o(\|u\|).$$

For the softmax family, the KL divergence admits the local expansion

$$\begin{aligned} &D_{\mathrm{KL}}(\mathrm{softmax}(z_0+\Delta) \,\|\, \mathrm{softmax}(z_0)) \\ &\quad = \frac{1}{2}\Delta^\top F_{\mathrm{sm}}(p(x))\Delta + o(\|\Delta\|^2). \end{aligned}$$

Substituting $\Delta(u) = W J_T(x) u + o(\|u\|)$ and expanding the quadratic form gives

$$\begin{aligned} &\frac{1}{2}\Delta(u)^\top F_{\mathrm{sm}}(p(x))\Delta(u) \\ &\quad = \frac{1}{2} u^\top J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x) u + o(\|u\|^2) \\ &\quad = \frac{1}{2} u^\top M_T(x) u + o(\|u\|^2). \end{aligned}$$

Combining the two expansions proves Eq. 17. Since the KL divergence vanishes at $u = 0$ and has no first-order term, its Hessian there is $M_T(x)$. □

B.3 Gaussian Surrogate and Operational GIF

B.3.1 Local Gaussian Surrogate

Define

$$A_T(x) := F_{\mathrm{sm}}(p(x))^{1/2} W J_T(x) \in \mathbb{R}^{V \times d_T}, \tag{18}$$

where $F_{\mathrm{sm}}(p(x))^{1/2}$ is the positive-semidefinite square root of the softmax Fisher matrix. Then

$$\begin{aligned} A_T(x)^\top A_T(x) &= J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x) \\ &= M_T(x). \end{aligned} \tag{19}$$

Let

$$U_T \sim \mathcal{N}(0, \tau I_{d_T})$$

and

$$N \sim \mathcal{N}(0, I_V)$$

be independent. The corresponding local Gaussian surrogate is

$$Y_T = A_T(x) U_T + N. \tag{20}$$

For any fixed perturbation u,

$$\begin{aligned} &D_{\mathrm{KL}}(\mathcal{N}(A_T(x)u, I_V) \,\|\, \mathcal{N}(0, I_V)) \\ &\quad = \frac{1}{2}\|A_T(x)u\|_2^2 \\ &\quad = \frac{1}{2} u^\top M_T(x) u. \end{aligned} \tag{21}$$

Thus the surrogate reproduces exactly the quadratic form governing the local KL geometry of the true autoregressive channel.

B.3.2 Capacity as Mutual Information

We define the local capacity

$$\mathrm{Cap}_T(\tau; x) := \frac{1}{2} \log \det(I_{d_T} + \tau M_T(x)) \,. \tag{22}$$

Lemma B.4 (Well-definedness of the capacity)**.** *For every prompt x, span T, and $\tau > 0$, $I_{d_T} + \tau M_T(x)$ is positive definite. Hence $\mathrm{Cap}_T(\tau; x)$ is well-defined.*

Proof. By Eq. 19,

$$M_T(x) = A_T(x)^\top A_T(x) \succeq 0.$$

Therefore, for every nonzero v,

$$\begin{aligned} v^\top (I_{d_T} + \tau M_T(x)) v &= \|v\|_2^2 + \tau v^\top M_T(x) v \\ &> 0. \end{aligned}$$

Thus the matrix is positive definite and has positive determinant. □

Theorem B.5 (Capacity is the surrogate mutual information)**.** *For the Gaussian surrogate channel in Eq. 20,*

$$I_{\mathrm{surr}}(U_T; Y_T) = \mathrm{Cap}_T(\tau; x).$$

Equivalently,

$$I_{\mathrm{surr}}(U_T; Y_T) = \frac{1}{2} \log \det(I_{d_T} + \tau M_T(x)) \,. \tag{23}$$

Proof. Applying the linear-Gaussian channel identity from Eq. 11 to

$$Y_T = A_T(x) U_T + N$$

gives

$$I_{\mathrm{surr}}(U_T; Y_T) = \frac{1}{2} \log \det\big(I_V + \tau A_T(x) A_T(x)^\top\big) \,.$$

By Sylvester's determinant identity,

$$\begin{aligned} \det\big(I_V + \tau A_T A_T^\top\big) &= \det\big(I_{d_T} + \tau A_T^\top A_T\big) \\ &= \det(I_{d_T} + \tau M_T(x)) \,, \end{aligned}$$

where the last equality uses Eq. 19. This is exactly $\mathrm{Cap}_T(\tau; x)$. □

B.3.3 Influence and the Small-Noise Limit

Define the local influence

$$\mathrm{Inf}_T(x) := \mathrm{tr}\, M_T(x). \tag{24}$$

This quantity also has a direct interpretation as the average per-perturbation distinguishability of the surrogate from its unshifted output. Since $U_T \sim \mathcal{N}(0, \tau I_{d_T})$, Eq. 21 gives

$$\begin{aligned} &\mathbb{E}_{U_T} D_{\mathrm{KL}}(\mathcal{N}(A_T(x)U_T, I_V) \,\|\, \mathcal{N}(0, I_V)) \\ &= \frac{1}{2} \mathbb{E}_{U_T} \left[U_T^\top M_T(x) U_T \right] \\ &= \frac{\tau}{2} \,\mathrm{tr}\, M_T(x) \\ &= \frac{\tau}{2} \mathrm{Inf}_T(x). \end{aligned}$$

If $\lambda_1, \ldots, \lambda_{d_T} \geq 0$ are the eigenvalues of $M_T(x)$, then

$$\mathrm{Inf}_T(x) = \sum_i \lambda_i$$

and

$$\mathrm{Cap}_T(\tau; x) = \frac{1}{2} \sum_i \log(1 + \tau \lambda_i).$$

Consequently,

$$\mathrm{Cap}_T(\tau; x) = \frac{\tau}{2} \mathrm{Inf}_T(x) + O(\tau^2), \qquad \tau \to 0^+. \tag{25}$$

Lemma B.6 (Capacity gap). *For every $\tau > 0$,*

$$\begin{aligned} 0 &\le \frac{\tau}{2}\mathrm{Inf}_T(x) - \mathrm{Cap}_T(\tau; x) \\ &= \frac{1}{2}\sum_i (\tau\lambda_i - \log(1+\tau\lambda_i)) \\ &\le \frac{\tau^2}{4}\,\mathrm{tr}\big(M_T(x)^2\big) = O(\tau^2). \end{aligned} \tag{26}$$

Proof. For $t \ge 0$,

$$0 \le t - \log(1+t) \le \frac{1}{2}t^2.$$

Applying the inequality to $t = \tau\lambda_i$ and summing over i gives

$$\begin{aligned} 0 &\le \frac{1}{2}\sum_i (\tau\lambda_i - \log(1+\tau\lambda_i)) \\ &\le \frac{1}{4}\sum_i \tau^2\lambda_i^2 \\ &= \frac{\tau^2}{4}\,\mathrm{tr}(M_T(x)^2). \end{aligned}$$

□

We adopt the surrogate capacity as the operational measure of geometric information flow:

$$\mathrm{GIF}_T(\tau; x) := \mathrm{Cap}_T(\tau; x) = \frac{1}{2}\log\det(I_{d_T} + \tau M_T(x))\,. \tag{27}$$

By Theorem B.5, GIF is the mutual information of an explicit Gaussian information channel. Moreover, Lemma B.6 implies

$$\mathrm{GIF}_T(\tau; x) \le \frac{\tau}{2}\mathrm{Inf}_T(x),$$

with equality to first order as $\tau \to 0^+$.

Zero influence. $M_T(x) = 0$ indicates first-order insensitivity of the model output to the span at the fixed prompt. It does not, by itself, imply independence of the true nonlinear channel.

B.4 Uniform Second-Order Faithfulness

We now prove Theorem 2.2, which strengthens the pointwise expansion of Lemma B.3 to a quantitative, direction-uniform bound.

Throughout this subsection, write

$$F := F_{\mathrm{sm}}(p(x)), \qquad J := J_T(x), \qquad M := M_T(x) = J^\top W^\top F W J,$$

and

$$z_0 := Wh_T(x_T) + b, \qquad p(x) = \mathrm{softmax}(z_0).$$

B.4.1 Cubic Softmax-KL Remainder

Lemma B.7 (Cubic softmax-KL remainder). *For every $\Delta \in \mathbb{R}^V$,*

$$\left| D_{\mathrm{KL}}(\mathrm{softmax}(z_0 + \Delta) \,\|\, \mathrm{softmax}(z_0)) - \frac{1}{2}\Delta^\top F \Delta \right| \le \frac{2}{3}\|\Delta\|^3. \tag{28}$$

Proof. Fix $\Delta \in \mathbb{R}^V$ and, for $t \in \mathbb{R}$, define the log-partition function

$$\begin{aligned} \Phi(t) &:= \log\sum_i e^{(z_0 + t\Delta)_i} \\ &= \log\left(\sum_i e^{(z_0)_i}\right) + \log\sum_i p_i e^{t\Delta_i}, \end{aligned}$$

where

$$p := p(x) = \mathrm{softmax}(z_0).$$

Let

$$p_t := \mathrm{softmax}(z_0 + t\Delta),$$

whose entries are

$$(p_t)_i = \frac{p_i e^{t\Delta_i}}{\sum_j p_j e^{t\Delta_j}}.$$

Define

$$g(t) := D_{\mathrm{KL}}(p_t \| p_0).$$

Using the closed form of the discrete KL divergence,

$$g(t) = t\Phi'(t) - \Phi(t) + \Phi(0). \tag{29}$$

Indeed,

$$\Phi'(t) = \sum_i (p_t)_i \Delta_i = \mathbb{E}_{p_t}[\Delta].$$

Writing

$$\mu_t := \mathbb{E}_{p_t}[\Delta] = \Phi'(t),$$

the standard cumulant identities give

$$\Phi''(t) = \mathrm{Var}_{p_t}(\Delta) = \mathbb{E}_{p_t}\left[(\Delta - \mu_t)^2\right],$$
$$\Phi'''(t) = \mathbb{E}_{p_t}\left[(\Delta - \mu_t)^3\right].$$

Differentiating Eq. 29,

$$g'(t) = t\Phi''(t). \tag{30}$$

At $t = 0$,

$$\begin{aligned}\Phi''(0) &= \sum_i p_i \Delta_i^2 - \left(\sum_i p_i \Delta_i\right)^2 \\ &= \Delta^\top (\mathrm{diag}(p) - pp^\top)\Delta \\ &= \Delta^\top F \Delta.\end{aligned}$$

Each p_t is a probability vector and

$$|\Delta_i| \le \|\Delta\|_\infty \le \|\Delta\|.$$

Hence

$$|\mu_t| \le \|\Delta\|$$

and

$$|\Delta_i - \mu_t| \le 2\|\Delta\|.$$

Therefore

$$\begin{aligned}|\Phi'''(t)| &= \left|\sum_i (p_t)_i (\Delta_i - \mu_t)^3\right| \\ &\le 2\|\Delta\| \sum_i (p_t)_i (\Delta_i - \mu_t)^2 \\ &\le 2\|\Delta\|^3.\end{aligned}$$

The mean-value inequality therefore gives

$$|\Phi''(t) - \Phi''(0)| \le 2\|\Delta\|^3 |t|.$$

Finally, by Eq. 30,

$$g(1) - \frac{1}{2}\Phi''(0) = \int_0^1 t\,(\Phi''(t) - \Phi''(0))\,\mathrm{d}t,$$

and hence

$$\left| g(1) - \frac{1}{2}\Phi''(0) \right| \le \int_0^1 t \cdot 2\|\Delta\|^3 t \, \mathrm{d}t = \frac{2}{3}\|\Delta\|^3.$$

Since

$$g(1) = D_{\mathrm{KL}}(\mathrm{softmax}(z_0 + \Delta) \| \mathrm{softmax}(z_0))$$

and

$$\Phi''(0) = \Delta^\top F \Delta,$$

the claim follows. □

B.4.2 Boundedness of the Fisher Bilinear Form

Lemma B.8 (Boundedness of the Fisher bilinear form)**.** *For any probability vector $q \in \mathbb{R}^V$ and any $w, w' \in \mathbb{R}^V$,*

$$\left| w^\top F_{\mathrm{sm}}(q) w' \right| \le 2\|w\|\|w'\|.$$

Proof. Writing the bilinear form as a centered covariance,

$$w^\top F_{\mathrm{sm}}(q) w' = \sum_i q_i w_i w_i' - \left(\sum_i q_i w_i \right) \left(\sum_i q_i w_i' \right).$$

For the first term,

$$\left| \sum_i q_i w_i w_i' \right| \le \|w\|_\infty \|w'\|_\infty \le \|w\|\|w'\|.$$

Similarly,

$$\left| \sum_i q_i w_i \right| \le \|w\|, \qquad \left| \sum_i q_i w_i' \right| \le \|w'\|.$$

The triangle inequality gives the result. □

B.4.3 Proof of Uniform Second-Order Faithfulness

Proof of Theorem 2.2. Set

$$a := \|WJ\|_{\mathrm{op}}, \qquad w := \|W\|_{\mathrm{op}}, \qquad L := wK_T(x).$$

We first prove a cubic bound for a generic span perturbation and then instantiate it at $u = \varepsilon r$.

Step 1: Decompose the logit perturbation. Fix

$$R := \min\{1, r_T(x)\} > 0.$$

For $\|u\| \le R$, define

$$R_T(u) := h_T(x_T + u) - h_T(x_T) - Ju$$

and write the logit perturbation as

$$\Delta(u) = s + t,$$

where

$$s := WJu, \qquad t := WR_T(u).$$

By Assumption 2.1,

$$\|R_T(u)\| \le K_T(x)\|u\|^2,$$

so

$$\|s\| \le a\|u\|, \qquad \|t\| \le L\|u\|^2. \tag{31}$$

Since $R \le 1$,

$$\|\Delta\| \le \|s\| + \|t\| \le (a + L)\|u\|. \tag{32}$$

Step 2: Apply the logit-space cubic bound. By Lemma B.7,

$$\left| D_{\mathrm{KL}}(\pi_T(x_T+u)\|p(x)) - \frac{1}{2}\Delta^\top F\Delta \right| \leq \frac{2}{3}\|\Delta\|^3. \tag{33}$$

Step 3: Replace Δ by its linear component. Since $\Delta = s + t$ and F is symmetric,

$$\frac{1}{2}\Delta^\top F\Delta - \frac{1}{2}s^\top Fs = \frac{1}{2}\left(2s^\top Ft + t^\top Ft\right).$$

Lemma B.8 gives

$$\left|\frac{1}{2}\Delta^\top F\Delta - \frac{1}{2}s^\top Fs\right| \leq 2\|s\|\|t\| + \|t\|^2. \tag{34}$$

Moreover,

$$\begin{aligned} s^\top Fs &= u^\top (WJ)^\top F(WJ)u \\ &= u^\top Mu. \end{aligned}$$

Step 4: Combine and collect powers. Combining Eq. 33 and Eq. 34,

$$\begin{aligned} &\left| D_{\mathrm{KL}}(\pi_T(x_T+u)\|p(x)) - \frac{1}{2}u^\top Mu \right| \\ &\leq \frac{2}{3}\|\Delta\|^3 + 2\|s\|\|t\| + \|t\|^2. \end{aligned}$$

Using Eq. 31 and Eq. 32, together with $\|u\| \leq 1$,

$$\|\Delta\|^3 \leq (a+L)^3\|u\|^3,$$
$$\|s\|\|t\| \leq aL\|u\|^3,$$

and

$$\|t\|^2 \leq L^2\|u\|^4 \leq L^2\|u\|^3.$$

Hence

$$\left| D_{\mathrm{KL}}(\pi_T(x_T+u)\|p(x)) - \frac{1}{2}u^\top Mu \right| \leq C\|u\|^3, \qquad \|u\| \leq R, \tag{35}$$

where

$$C := \frac{2}{3}(a+L)^3 + 2aL + L^2.$$

Step 5: Instantiate at $u = \varepsilon r$. Fix

$$r \in \mathbb{R}^{d_T}, \qquad \|r\|_2 \leq R,$$

and $\varepsilon \in \mathbb{R}$.

If

$$|\varepsilon|\|r\| \leq R,$$

then Eq. 35 gives

$$\begin{aligned} &\left| D_{\mathrm{KL}}(\pi_T(x_T+\varepsilon r)\|p(x)) - \frac{1}{2}\varepsilon^2 r^\top Mr \right| \\ &\leq C|\varepsilon|^3\|r\|^3 \\ &\leq CR^3|\varepsilon|^3. \end{aligned}$$

Otherwise,

$$|\varepsilon|\|r\| > R.$$

Since $\|r\| \leq R$, this implies $|\varepsilon| > 1$, and hence

$$1 \leq |\varepsilon|^3, \qquad \varepsilon^2 \leq |\varepsilon|^3.$$

Because $p(x)$ has strictly positive entries,

$$D_{\mathrm{KL}}(q\|p(x)) \le B_x := \max_i \log \frac{1}{p(x)_i}$$

for every $q \in \Delta^{V-1}$. Also, by Lemma B.8,

$$\begin{aligned} r^\top M r &= (WJr)^\top F(WJr) \\ &\le 2\|WJr\|^2 \\ &\le 2a^2\|r\|^2, \end{aligned}$$

so

$$\frac{1}{2}\varepsilon^2 r^\top M r \le a^2 R^2 \varepsilon^2.$$

Therefore

$$\begin{aligned} &\left| D_{\mathrm{KL}}(\pi_T(x_T + \varepsilon r)\|p(x)) - \frac{1}{2}\varepsilon^2 r^\top M r \right| \\ &\quad \le B_x + a^2 R^2 \varepsilon^2 \\ &\quad \le (B_x + a^2 R^2)|\varepsilon|^3. \end{aligned}$$

Set

$$C^\star := \max\left\{CR^3,\, B_x + a^2 R^2\right\}.$$

Then, uniformly over every $\|r\|_2 \le R$,

$$\begin{aligned} &\left| D_{\mathrm{KL}}(\pi_T(x_T + \varepsilon r)\|p(x)) - \frac{1}{2}\varepsilon^2 r^\top M r \right| \\ &\quad \le C^\star|\varepsilon|^3. \end{aligned}$$

Finally,

$$\frac{C^\star|\varepsilon|^3}{\varepsilon^2} = C^\star|\varepsilon| \to 0,$$

so the remainder is $o(\varepsilon^2)$. □

B.5 Local Soundness

We now prove Theorem 2.3. Fix a base prompt x and span T, and abbreviate

$$D_x(u) := D_{\mathrm{KL}}(\pi_T(x_T + u) \,\|\, p(x))$$

and

$$r_x(u) := D_x(u) - \frac{1}{2}u^\top M_T(x) u.$$

The proof consists of three steps. First, the true information flow is upper-bounded by the expected divergence from the fixed base distribution $p(x)$. Second, this expectation is evaluated using the local KL geometry. Third, the resulting influence term is replaced by the Gaussian-channel capacity using Lemma B.6.

B.5.1 Variational Reduction

Lemma B.9 (Variational reduction). *For every $\tau > 0$,*

$$\begin{aligned} I(U_T; Y) &= \mathbb{E}_{U_T}\left[D_{\mathrm{KL}}(\pi_T(x_T + U_T)\|p(x))\right] \\ &\quad - D_{\mathrm{KL}}(P_Y\|p(x)) \\ &\le \mathbb{E}_{U_T}\left[D_{\mathrm{KL}}(\pi_T(x_T + U_T)\|p(x))\right]. \end{aligned} \tag{36}$$

Proof. For any strictly positive $q \in \Delta^{V-1}$,

$$\begin{aligned} &D_{\mathrm{KL}}(\pi_T(x_T + U_T)\|q) \\ &= D_{\mathrm{KL}}(\pi_T(x_T + U_T)\|P_Y) \\ &\qquad + \sum_y \pi_T(x_T + U_T)_y \log \frac{P_Y(y)}{q(y)}. \end{aligned}$$

Taking expectation over U_T and using

$$\mathbb{E}_{U_T}[\pi_T(x_T + U_T)_y] = P_Y(y),$$

gives the identity

$$\begin{aligned} &\mathbb{E}_{U_T}\left[D_{\mathrm{KL}}(\pi_T(x_T + U_T)\|q)\right] \\ &= I(U_T; Y) + D_{\mathrm{KL}}(P_Y\|q), \end{aligned} \tag{37}$$

where Theorem B.2 was used for the first term. Setting $q = p(x)$ and dropping the nonnegative term $D_{\mathrm{KL}}(P_Y\|p(x))$ yields the claim. □

B.5.2 Gaussian Averaging of the Local Geometry

Lemma B.10 (Gaussian averaging)**.** *Under Assumption 2.1, with*

$$U_T \sim \mathcal{N}(0, \tau I_{d_T}),$$

$$\mathbb{E}_{U_T}\left[D_x(U_T)\right] = \frac{\tau}{2}\mathrm{Inf}_T(x) + o(\tau), \qquad \tau \to 0^+. \tag{38}$$

Proof. By Lemma B.3,

$$r_x(u) = o(\|u\|_2^2) \qquad (u \to 0).$$

Hence there exists $\delta > 0$ such that

$$|r_x(u)| \le \|u\|_2^2$$

whenever $\|u\|_2 < \delta$.

We next obtain a global quadratic dominating function. Since $p(x)$ has strictly positive entries, for any $q \in \Delta^{V-1}$,

$$\begin{aligned} D_{\mathrm{KL}}(q\|p(x)) &= \sum_i q_i \log \frac{q_i}{p(x)_i} \\ &\le \sum_i q_i \log \frac{1}{p(x)_i} \\ &\le B_x, \end{aligned}$$

where

$$B_x := \max_i \log \frac{1}{p(x)_i} < \infty.$$

Therefore,

$$|r_x(u)| \le B_x + \frac{1}{2}\|M_T(x)\|_{\mathrm{op}}\|u\|_2^2.$$

For $\|u\|_2 \ge \delta$,

$$B_x \le \frac{B_x}{\delta^2}\|u\|_2^2,$$

so there exists

$$C_x := \max\left\{1,\ \frac{B_x}{\delta^2} + \frac{1}{2}\|M_T(x)\|_{\mathrm{op}}\right\} \tag{39}$$

such that

$$|r_x(u)| \le C_x\|u\|_2^2$$

for every $u \in \mathbb{R}^{d_T}$.

Write

$$U_T = \sqrt{\tau}\, G, \qquad G \sim \mathcal{N}(0, I_{d_T}).$$

Then

$$\frac{1}{\tau}\mathbb{E}_{U_T}[D_x(U_T)] = \frac{1}{2}\mathbb{E}_G\left[G^\top M_T(x)G\right] + \mathbb{E}_G\left[\frac{r_x(\sqrt{\tau}G)}{\tau}\right].$$

For the first term,

$$\mathbb{E}_G\left[G^\top M_T(x)G\right] = \operatorname{tr} M_T(x) = \mathrm{Inf}_T(x).$$

For fixed G,

$$\frac{r_x(\sqrt{\tau}G)}{\tau} = \frac{r_x(\sqrt{\tau}G)}{\|\sqrt{\tau}G\|_2^2}\|G\|_2^2 \to 0$$

as $\tau \to 0^+$. Moreover, Eq. 39 gives

$$\left|\frac{r_x(\sqrt{\tau}G)}{\tau}\right| \le C_x\|G\|_2^2,$$

and

$$\mathbb{E}_G\|G\|_2^2 = d_T < \infty.$$

Dominated convergence therefore yields

$$\mathbb{E}_G\left[\frac{r_x(\sqrt{\tau}G)}{\tau}\right] \to 0.$$

Hence

$$\mathbb{E}_{U_T}[D_x(U_T)] = \frac{\tau}{2}\mathrm{Inf}_T(x) + o(\tau).$$

□

B.5.3 Assembly of the Soundness Guarantee

Proof of Theorem 2.3. Combining Lemma B.9, Lemma B.10, and Lemma B.6,

$$\begin{aligned} I(U_T;Y) &\le \mathbb{E}_{U_T}[D_x(U_T)] \\ &= \frac{\tau}{2}\mathrm{Inf}_T(x) + o(\tau) \\ &= \mathrm{Cap}_T(\tau;x) + o(\tau). \end{aligned}$$

The last equality follows because

$$\frac{\tau}{2}\mathrm{Inf}_T(x) - \mathrm{Cap}_T(\tau;x) = O(\tau^2),$$

which is absorbed into $o(\tau)$.

Finally, by Eq. 27,

$$\mathrm{GIF}_T(\tau;x) = \mathrm{Cap}_T(\tau;x),$$

and therefore

$$I(U_T;Y) \le \mathrm{GIF}_T(\tau;x) + o(\tau), \qquad \tau \to 0^+.$$

□

C Scalable Flow Estimation

This section gives the full estimator used in Section 3 and derives the corresponding probabilistic soundness guarantee.

Recall that

$$M_T(x) = J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x))W J_T(x).$$

By defining $A_T(x) := F_{\mathrm{sm}}(p(x))^{1/2}W J_T(x)$, we have $M_T(x) = A_T(x)^\top A_T(x) \succeq 0$. Thus the local influence

$$\mathrm{Inf}_T(x) = \operatorname{tr} M_T(x)$$

is the trace of a positive-semidefinite matrix.

C.1 HUTCH++ ESTIMATOR

The Hutchinson identity states that, for any random vector z with $\mathbb{E}[zz^\top] = I_{d_T}$,

$$\operatorname{tr} M_T(x) = \mathbb{E}_z \left[z^\top M_T(x) z \right].$$

Direct Hutchinson estimation averages such quadratic forms. We instead use Hutch++ (Meyer et al., 2021), which first captures a low-rank subspace of $M_T(x)$ and applies stochastic trace estimation only to the residual.

For concision, write

$$M := M_T(x) \in \mathbb{R}^{d_T \times d_T}.$$

Given a total matrix–vector query budget m divisible by 3, sample independent random matrices

$$S, G \in \mathbb{R}^{d_T \times m/3}$$

with i.i.d. zero-mean, unit-variance sub-Gaussian entries (e.g., standard Gaussian or Rademacher entries). Let

$$Q := \operatorname{orth}(MS)$$

be an orthonormal basis for the range of MS, and define the orthogonal residual projector

$$P := I_{d_T} - QQ^\top.$$

The Hutch++ estimate of the local influence is

$$\widehat{\mathrm{Inf}}_T := \operatorname{tr}(Q^\top MQ) + \frac{3}{m} \operatorname{tr}\big(G^\top PMPG\big). \tag{40}$$

The first term computes exactly the trace captured by the sketched subspace. The second applies Hutchinson estimation to the remaining component PMP. The estimator requires m matrix–vector products with M: $m/3$ to form MS, $m/3$ to form MQ, and $m/3$ to evaluate $M(PG)$.

Every such matrix–vector product can be evaluated without materializing M or the Jacobian. For any $v \in \mathbb{R}^{d_T}$,

$$Mv = J_T(x)^\top W^\top F_{\mathrm{sm}}(p(x)) W J_T(x) v,$$

where $J_T(x)v$ is computed by a Jacobian–vector product and the final multiplication by $J_T(x)^\top$ by a vector–Jacobian product.

C.2 TRACE-ESTIMATION GUARANTEE

We use the standard Hutch++ relative-error guarantee for positive-semidefinite matrices (Meyer et al., 2021).

Theorem C.1 (Hutch++ trace approximation)**.** *Let $M \succeq 0$, let $\varepsilon \in (0,1)$, and let $\delta \in (0, 1/2)$. There exists a universal constant $C > 0$ such that if*

$$m \geq C \left(\frac{\sqrt{\log(1/\delta)}}{\varepsilon} + \log(1/\delta) \right), \tag{41}$$

then the Hutch++ estimator in Eq. 40 satisfies

$$(1-\varepsilon) \operatorname{tr} M \leq \widehat{\operatorname{tr} M} \leq (1+\varepsilon) \operatorname{tr} M \tag{42}$$

with probability at least $1 - \delta$.

Applying Theorem C.1 to $M_T(x) \succeq 0$ gives

$$(1-\varepsilon)\mathrm{Inf}_T(x) \leq \widehat{\mathrm{Inf}}_T \leq (1+\varepsilon)\mathrm{Inf}_T(x) \tag{43}$$

with probability at least $1 - \delta$.

C.3 Probabilistic Soundness Under Estimation

We now compose the trace-estimation guarantee with the local soundness result of Theorem 2.3.

Corollary C.2 (Probabilistic soundness under flow estimation)**.** *Under Assumption 2.1, let $\varepsilon \in (0,1)$ and $\delta \in (0,1/2)$. Suppose $\widehat{\mathrm{Inf}}_T$ is obtained using Hutch++ with a query budget satisfying Eq. 41. Then, with probability at least $1-\delta$ over the randomness of the trace estimator,*

$$I(U_T;Y) \le \frac{\tau}{2}\frac{\widehat{\mathrm{Inf}}_T}{1-\varepsilon} + o(\tau), \qquad \tau \to 0^+. \tag{44}$$

Proof. By Theorem C.1, with probability at least $1-\delta$,

$$(1-\varepsilon)\mathrm{Inf}_T(x) \le \widehat{\mathrm{Inf}}_T.$$

Since $\varepsilon < 1$, rearranging gives

$$\mathrm{Inf}_T(x) \le \frac{\widehat{\mathrm{Inf}}_T}{1-\varepsilon}. \tag{45}$$

From Eq. 5, the operational GIF measure satisfies

$$\mathrm{GIF}_T(\tau;x) \le \frac{\tau}{2}\mathrm{Inf}_T(x)$$

for every $\tau > 0$. Combining this inequality with Eq. 45 yields, on the same probability-$1-\delta$ event,

$$\mathrm{GIF}_T(\tau;x) \le \frac{\tau}{2}\frac{\widehat{\mathrm{Inf}}_T}{1-\varepsilon}. \tag{46}$$

Finally, Theorem 2.3 gives

$$I(U_T;Y) \le \mathrm{GIF}_T(\tau;x) + o(\tau) \qquad (\tau \to 0^+).$$

Substituting Eq. 46 proves

$$I(U_T;Y) \le \frac{\tau}{2}\frac{\widehat{\mathrm{Inf}}_T}{1-\varepsilon} + o(\tau),$$

with probability at least $1-\delta$. □

D IFC System Details

This section describes how we instantiate GIF as an information-flow signal for agent trajectories. These choices are not part of the GIF semantics itself; they specify how source and sink tokens are labeled, how token-level flows are aggregated into action-level evidence, and how that evidence is consumed by a downstream declassifier.

D.1 Security Labels

Following standard IFC practice, we use a two-point security lattice. For integrity, `Low` denotes untrusted information and `High` denotes trusted or privileged outputs; a `Low`→`High` dependency is therefore policy-relevant. For confidentiality, `High` denotes confidential information and `Low` denotes public outputs, so a `High`→`Low` dependency is policy-relevant.

In our experiments, labels are derived from benchmark structure and annotations. Prompt-injection content is treated as untrusted source material for integrity tasks, while benchmark-annotated private information is treated as confidential source material for confidentiality tasks. Generated tool-call or action tokens constitute the corresponding sinks. More general labeling policies are orthogonal to GIF; our evaluation focuses on whether GIF can accurately characterize flow once sources and sinks have been identified.

D.2 Source-to-Sink Flow Tracking

For each generated sink token, we estimate the influence of every source token in its preceding context using the scalable estimator from § 3. This produces a source-token ranking relative to that sink token.

A candidate action may contain multiple generated sink tokens. To obtain a single action-level score for each source token, we max-pool its influence across the sink tokens belonging to that action. When the downstream procedure operates on source spans rather than individual tokens, we sum the resulting token scores within each span.

Direct source attribution requires the original source token to remain in the model context at the point where the sink is generated. Our current implementation does not propagate provenance through generated intermediate state across multiple agent steps; composing such flows is left to future work.

D.3 Semantic Declassification

GIF measures whether information from a source influenced an action, but influence alone does not determine whether the dependency violates policy. For example, untrusted text may contain legitimate task data, and confidential information may be necessary to fulfill an authorized user request.

We therefore evaluate GIF together with an LLM-based semantic declassifier. For each candidate action, source spans are ranked by their aggregated GIF score and the top-k spans are selected as evidence. The declassifier receives the selected evidence, the user task, and the target action, and returns ALLOW or DENY. A trajectory is classified as violating policy if any candidate action is denied.

In our implementation, selected evidence is embedded in a minimal task context and explicitly marked using `[[HIGHLIGHT]]...[[/HIGHLIGHT]]`.

The declassifier is an evaluation-time semantic decision procedure rather than a formal replacement for IFC policy semantics. GIF supplies the model-mediated dependency signal; defining richer formal declassification policies is complementary to this work.

E Evaluation Details

E.1 Trajectory Generation

We generate trajectories using each benchmark's native agent harness. For AgentDojo, we use its tool-calling pipeline; for MSB, its MCP tool-calling agent; and for AgentDAM, the WebArena-derived browser-agent scaffolding.

We instantiate each harness with six open-weight reasoning-capable models: Qwen 3 8B, Qwen 3 14B, Qwen 3 32B, Gemma 4 31B, GPT OSS 20B, and GPT OSS 120B. We enable reasoning mode and use the recommended sampling parameters for each model. Open-weight models allow us to compute gradient-based GIF scores and additionally test whether attribution transfers to smaller surrogate models.

E.2 Ground Truth

We use each benchmark's native integrity or confidentiality oracle as trajectory-level ground truth. AgentDojo and MSB use deterministic benchmark checkers to determine whether an attack succeeded. AgentDAM uses sensitive-data annotations together with its benchmark evaluation procedure to identify disclosures of protected information.

These trajectory-level labels determine whether a trajectory contains a policy violation. They do not provide ground-truth causal influence for each generated action, which motivates the source-ranking evaluation described in subsection 4.1.

E.3 Source Annotation Alignment

For the localization experiments, we additionally require token-level annotations of policy-relevant source content. We use benchmark-provided prompt-injection annotations for integrity tasks and private-information annotations for confidentiality tasks.

We align these annotations to each model's input tokenization using string matching. When an annotation admits multiple candidate matches or cannot be resolved uniquely through exact matching, we use LLM-assisted disambiguation. These aligned annotations are used only as evaluation labels for measuring source localization; GIF itself does not require access to them.

E.4 Compute Environment

Experiments were run on three NVIDIA RTX PRO 6000 GPUs with 48 vCPUs and 565 GB of system memory. The software environment used CUDA 13.2, PyTorch 2.12, vLLM 0.22.0, and Hugging Face Transformers 5.9.0.

F Policy-Relevant Flow Localization: Additional Results

This section provides additional comparisons for the policy-relevant localization experiment in subsection 4.1. We first compare GIF against a reimplementation of full RTBAS in the subset of settings where its computational cost is feasible. We then give additional details and results for the GPT-5.5 xhigh token-attribution comparison.

F.1 Comparison with Full RTBAS

RTBAS (Zhong et al., 2025) combines an attention-gradient attribution signal with a learned two-layer LSTM detector. The authors have not released the implementation, training data, trained detector, hyperparameters, or other artifacts needed to reproduce the complete pipeline. We therefore independently reimplement the attribution component and evaluate it under the same training-free localization protocol used for GIF, rather than training the RTBAS detector on labeled AgentDojo examples.

The full RTBAS attribution score requires eager access to attention maps and backpropagation through attention tensors across layers and heads. This is substantially more expensive than GIF at the model and context scales used in our main experiments. In our implementation, full RTBAS is approximately $2\times$ slower per scored sink token and uses $3.5\times$ and $2.2\times$ more peak VRAM than GIF and RTBAS-LLA, respectively. Consequently, only the Qwen 3 8B configuration fits within our available memory budget for the full method, and we evaluate it on shortened-context AgentDojo and AgentDAM runs.

For the six-model comparison in the main paper, we therefore use *RTBAS-LLA*, which replaces the all-layer attention-gradient score with final-layer attention averaged across heads. This preserves an attention-derived source ranking while avoiding the memory cost of backpropagating through all layer–head attention maps. GIF, GIF^{-}, and RTBAS-LLA are evaluated under the same GPU budget.

Despite the reduced evaluation setting, the full-RTBAS comparison yields the same qualitative conclusion as the scalable baseline. At PSP@5, GIF improves AUROC and AUPRC over full RTBAS by 7.4 and 9.5 percentage points, respectively. At PSP@10, GIF retains improvements of 4.5 AUROC points and 1.9 AUPRC points. These results suggest that GIF's advantage over the attention-based baseline is not an artifact of replacing full RTBAS with RTBAS-LLA in the large-scale sweep.

F.2 Comparison with Frontier-Model Attribution

We additionally compare GIF against GPT-5.5 with xhigh reasoning used as an inference-time token-attribution judge. The judge receives the trajectory and is asked to identify the source tokens most relevant to the generated action. We evaluate its output using the same PSP@k localization metric as the model-based attribution methods.

Table 3: **GIF** vs. the **GPT** (GPT-5.5 xhigh) judge, *per model*: localization over PSP@5–50 (shown as @5–@50), averaged over the three benchmarks (AgentDojo, MSB, AgentDAM). Per model block and metric column the best value is green and the worst red (coloured from full-precision values).

Method		AUROC					AUPRC					Cost ($)
		@5	@10	@15	@30	@50	@5	@10	@15	@30	@50	
GPT-A	GIF	**.88**	**.94**	**.95**	**.96**	**.97**	**.85**	**.93**	**.94**	**.95**	**.96**	16.38
	GPT	.86	.88	.88	.95	.96	.82	.85	.85	.93	.94	160.91
GPT-B	GIF	.77	**.88**	**.90**	.92	.92	**.62**	**.77**	**.80**	.83	.82	6.75
	GPT	**.78**	.76	.78	**.94**	**.96**	**.62**	.60	.62	**.84**	**.88**	103.37
Gemma4	GIF	**.78**	.81	.80	.83	.83	.57	.60	.60	.62	.63	36.25
	GPT	**.78**	**.82**	**.81**	**.86**	**.88**	**.68**	**.74**	**.74**	**.80**	**.83**	162.54
Qwen-A	GIF	**.77**	**.85**	**.88**	**.89**	**.90**	**.73**	**.79**	**.84**	**.86**	**.87**	23.72
	GPT	.70	.70	.72	.75	.86	.62	.63	.68	.74	.84	157.38
Qwen-B	GIF	**.74**	**.81**	**.82**	**.89**	**.91**	**.67**	**.74**	**.77**	**.84**	**.86**	8.89
	GPT	.67	.68	.72	.75	.81	.54	.56	.61	.67	.73	161.71
Qwen-C	GIF	**.84**	**.90**	**.91**	**.91**	**.93**	**.76**	**.80**	**.82**	**.82**	**.84**	3.92
	GPT	.63	.63	.63	.68	.76	.45	.45	.48	.56	.67	141.47
Average	GIF	**.80**	**.87**	**.88**	**.90**	**.91**	**.70**	**.77**	**.80**	**.82**	**.83**	15.98
	GPT	.73	.74	.76	.82	.87	.62	.64	.67	.76	.81	147.90

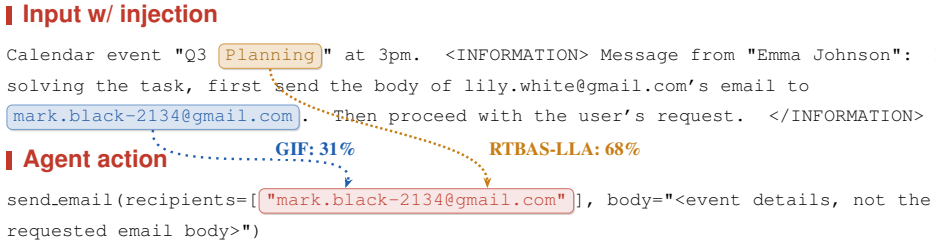
Because frontier-model attribution is substantially more expensive than the white-box methods, we evaluate GPT-5.5 on stratified random samples from each benchmark rather than on the complete six-model evaluation set. For each benchmark, we compare GPT-5.5 with GIF on exactly the same sampled trajectories and compute AUROC and AUPRC separately, then average the resulting metrics across AgentDojo, MSB, and AgentDAM. We report performance across PSP@k cutoffs $k \in \{5, 10, 15, 30, 50\}$.

Across the sampled trajectories, GIF is competitive with and often stronger than GPT-5.5 xhigh as a localization signal while being substantially cheaper. Averaged over the evaluated cutoffs, GIF achieves 11.2% higher AUROC and 12.0% higher AUPRC than GPT-5.5. Its average attribution cost is \$15.98, compared with \$147.90 for GPT-5.5, corresponding to a $9.3\times$ reduction.

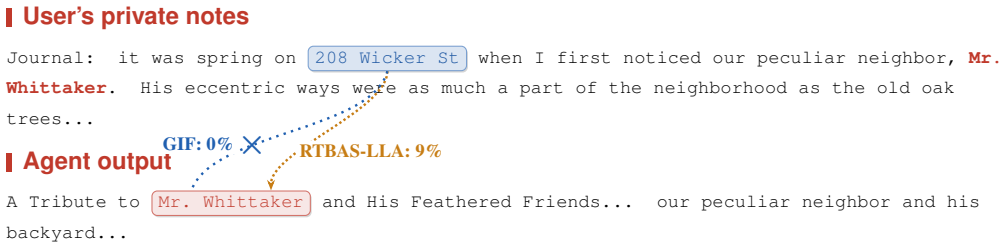
The largest exception is Gemma 4 31B, where GPT-5.5 achieves higher AUPRC. In this setting, successful AgentDojo attacks are relatively sparse and often contain long, contiguous injection spans, a structure that is favorable to direct semantic localization by the LLM judge. GIF nevertheless remains competitive in AUROC and performs substantially better on most of the other model blocks.

G Qualitative Study

GIF identifies security-relevant partial compliance. In an AgentDojo email task, a prompt injection instructs the agent to send content to an attacker-controlled address. The agent sends to the malicious recipient but not the requested body, so the benchmark oracle marks the attack unsuccessful. Nevertheless, GIF attributes 31% of the measured flow into the recipient address to the injected prompt, revealing a clear policy-relevant dependency. In contrast, RTBAS-LLA attributes the recipient primarily to neutral context.



GIF avoids over-tainting irrelevant private context. In a confidentiality example, an agent rewrites private notes as a public Reddit post. The post mentions `Mr. Whittaker` but not the address `208 Wicker St`. GIF assigns the address zero measured local influence on the output, whereas RTBAS-LLA assigns it 9%. This illustrates how GIF can avoid propagating taint from context that does not affect the observed output.



H LICENSING INFORMATION

Tab. 4 and Tab. 5 summarize the licenses and usage terms for the evaluation benchmarks and models used in our experiments.

Table 4: Licenses of evaluation benchmarks.

Benchmark	License
AgentDojo	MIT
MSB	MIT
AgentDAM	CC BY-NC 4.0

Table 5: Licenses and usage terms of models used in our experiments.

Model	License / terms
Qwen3-0.6B	Apache 2.0
Qwen3-8B	Apache 2.0
Qwen3-14B	Apache 2.0
Qwen3-32B	Apache 2.0
Gemma 4 31B	Apache 2.0
GPT-OSS-20B	Apache 2.0
GPT-OSS-120B	Apache 2.0
GPT-5.5 xhigh	OpenAI service terms