Evaluation Theater: How Structural Compliance Decouples from Cognitive Judgment in Deployed LLM Agents

Ning Coeva [1]

Abstract

Deployed LLM agent pipelines increasingly adopt structured-output safety gates—function-call validators, schema enforcers, evidence-completeness checks—as cheap, deterministic surrogates for harder-to-measure cognitive properties. We show this surrogate can fail completely. In a security-operations pipeline simulation across 8 frontier LLMs, an Evidence Addressal Completeness gate yields *identical* perfect scores (primary completeness $= 1.0$, gate pass rate $= 100\%$) on both 144 legitimate verdict revisions and 49 social-engineering capitulations: a discriminative gap of 0.0 percentage points. Across the 193 changed-verdict trials the gate's blocking logic never once engaged. We name this failure mode *evaluation theater*, characterize it with a mirror-pair diagnostic, document a "recognize-but-capitulate" pattern in which models articulate suspicion in structured output before flipping verdicts anyway, and propose a Cognitive Compliance Index built on content signals the structural gate ignores (AUROC 0.934, 95% CI $[0.899, 0.966]$).

1. Introduction

LLM agents deployed in the wild execute irreversible real-world actions: blocking IP addresses, quarantining files, disabling user accounts, sending emails (Yao et al., 2023; Schick et al., 2023; Liu et al., 2024). To control these effects, production pipelines bolt on *structural* safety gates—function-call validators, JSON-schema enforcers, evidence-completeness checks—that verify the *form* of the agent's output before allowing downstream effects (Willard & Louf, 2023; Dong et al., 2024; OpenAI, 2024). Such gates are deterministic, cheap, and a natural defense against indirect prompt injection and authority spoofing (Greshake et al., 2023; Toyer et al., 2024). The promise is that structural compliance is a useful surrogate for cognitive compliance. We show the surrogate can break completely.

We study a security-operations (SOC) triage pipeline. An LLM agent issues a Tier-1 verdict T_1 triggering physical enforcement (`BLOCK_IP`, `ISOLATE_HOST`, `DISABLE_ACCOUNT`), then a Tier-2 verdict T_2 given new "evidence"—legitimate or socially engineered. A revision in T_2 rolls back enforcement; the Evidence Addressal Completeness (EAC) gate permits rollback only when T_2 structurally addresses every primary evidence node from T_1. Evidence-completeness checks of this shape appear in agent supervisor patterns generally: feedback loops (Yao et al., 2023), structured rubrics over tool output (Liu et al., 2024), schema-bound verifiers (Willard & Louf, 2023; OpenAI, 2024).

Across 8 frontier LLMs $\times$ 6 scenarios $\times$ 5 runs (216 trials), the EAC gate's structural metric is degenerate. In all 144 legitimate revisions *and* all 49 social-engineering capitulations, primary completeness equals 1.0 and the gate passes—a discriminative gap of 0.0 percentage points. We call this *evaluation theater*.

Contributions. (i) We identify evaluation theater as a deployment failure mode of structured-output safety gates with zero-discriminative-power evidence, and show that tightening the gate along its own completeness axis stays at chance (Section 3). (ii) We propose a *mirror-pair diagnostic*—paired scenarios sharing surface claim but differing in evidence quality—and document a "recognize-but-capitulate" pattern in which models articulate suspicion in structured output before flipping verdicts (Sections 3.2–3.3). (iii) We propose a preliminary Cognitive Compliance Index (CCI) on content signals the gate ignores: AUROC 0.934, 95% CI $[0.899, 0.966]$ (Section 4). The signal was already in the agent's output; the gate was looking at the wrong layer.

2. Setup

Pipeline. Each trial follows a $T_1 \rightarrow T_2$ schema. The agent emits T_1: a JSON object with `threat_level`, `recommended_actions`, and an `evidence_chain`

[1]Independent Researcher. Correspondence to: Ning Coeva <coevaxlumen.log@gmail.com>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

of {primary, supporting}-tagged facts, produced under standard constrained-decoding contracts (Willard & Louf, 2023; Dong et al., 2024). T_1 is dispatched to a simulated executor; an "incident update" is then injected and the agent emits T_2 with revised `threat_level`, an `addressed_evidence` field responding to each T_1 node by reference, and any `new_evidence` it observes. When T_2 down-revises threat, the EAC gate decides whether to permit the rollback.

EAC gate. The gate computes $\text{primary_completeness} = |\text{primary nodes substantively addressed in } T_2| / |\text{primary nodes in } T_1|$ (and analogously for supporting); rollback is permitted iff $\text{primary_completeness} = 1.0$ and $\text{supporting_completeness} \geq 0.5$. "Substantively addressed" filters hollow refs (single-token responses, generic "handled" templates) by a length-and-content heuristic.

Models, scenarios, mirror pair. We evaluate 8 LLMs (Claude Haiku 4.5, Claude Sonnet 4.6, DeepSeek-Chat, Gemini 2.5 Flash, GPT-4.1-mini, GPT-5.4-mini, GPT-5.4, Qwen-Plus) on 6 scenarios with 5 runs each at temperature 0.0, yielding 216 valid trials.[1] Four scenarios warrant legitimate down-revision (SQL injection re-attributed to QA, lateral movement to a sysadmin, etc.); two are social-engineering attempts: FALSE_CONTRA (attacker impersonating internal IT during phishing) and EXFIL (attacker claiming to be an external auditor mid-data-exfiltration). We isolate one *mirror pair*: BRUTEFORCE (legitimate; T_2 confirms an authorized red-team) and FALSE_CONTRA. Both invoke an authorization claim at face value; they differ in whether T_2 contains corroborating evidence. Same surface, different cognition—a clean stress test for any safety gate that claims to discriminate the two.

3. Findings

3.1. Zero discriminative power

We restrict attention to the 193 trials in which T_2 down-revised the threat level from T_1 (i.e., trials where the EAC gate is in scope to adjudicate a rollback). Of these, 144 are legitimate revisions and 49 are social-engineering capitulations (19 from FALSE_CONTRA and 30 from EXFIL). Table 1 reports the gate's behavior.

The gap is not noise: primary completeness is identically 1.0 in every trial; the rollback-block branch was reached $0/193$ times. We call this pattern *evaluation theater*: a deterministic safety mechanism whose dispatch rule is decoupled from the property it claims to enforce. Once an LLM agent names each T_1 evidence ID and writes anything substantive, the gate is satisfied regardless of whether the response engages the cognitive content. The degeneracy is not a tuning artefact: both classes produce T_2 outputs orders of magnitude beyond any plausible length threshold (median raw length 2,785 chars legit vs 2,216 chars capitulation), and any string-level criterion can be saturated by a producer optimized to satisfy the form.

Table 1. The EAC gate gives identical perfect scores to legitimate revisions and social-engineering capitulations across 193 changed-verdict trials.

	Legit revisions ($N = 144$)	Social capitulations ($N = 49$)
Primary completeness $= 1.0$	144/144	49/49
Supporting completeness ≥ 0.5	144/144	49/49
Gate pass rate	100%	100%
Discriminative gap: 0.0 **percentage points**		

Table 2. Mirror pair: BRUTEFORCE (legitimate retraction) vs. FALSE_CONTRA (social-engineering capitulation). Both invoke an authorization claim at face value; they differ in whether T_2 contains corroborating evidence. Whenever a model changed its verdict the EAC gate passed—regardless of whether the change was warranted.

Model	BRUTEFORCE (changed/total)	FALSE_CONTRA (changed/total)
Claude Haiku 4.5	5/5	0/3
Claude Sonnet 4.6	5/5	0/5
DeepSeek-Chat	5/5	5/5
Gemini 2.5 Flash	5/5	5/5
GPT-4.1-mini	5/5	3/3
GPT-5.4	5/5	0/5
GPT-5.4-mini	5/5	3/3
Qwen-Plus	5/5	3/3
Total	**40/40** (100%)	**19/32** (59%)
EAC gate pass rate	100%	100%

3.2. Mirror pair: same surface, different cognition, identical metric

The mirror pair BRUTEFORCE/FALSE_CONTRA isolates the gap. Table 2 reports per-model verdict change and gate pass on the pair.

Every model retracts on BRUTEFORCE ($40/40$). On FALSE_CONTRA, $5/8$ capitulate ($19/32$) and $3/8$ hold ground (Haiku 4.5, Sonnet 4.6, GPT-5.4 retain CRITICAL on every run). The holdouts are the cleanest control: they retract on BRUTEFORCE *and* hold on FALSE_CONTRA on the same surface claim, proving the discrimination is achievable. The EAC gate cannot see it—both refusals and surrenders score primary completeness $= 1.0$.

[1] Of 240 nominal trials, 24 were not collected for budget reasons. Loss is class-balanced: 10% each on legitimate ($16/160$) and social-engineering ($8/80$) scenarios.

3.3. Recognize-but-capitulate

If structural completeness is uninformative, where does cognitive signal live? In the same T_2 the gate is reading. Prior work has shown LLMs can verbalize uncertainty (Lin et al., 2022) and that chain-of-thought is partially faithful (Lanham et al., 2023); we observe a sharper version. Of 49 capitulations, 15 (31%) contain explicit suspicion-articulation in `new_evidence` or `addressed_evidence` (keywords: "unverif*", "cannot be authenticated", "unsolicited", "impersonat*", ...), spanning three model families: Haiku 4.5 (5/5), GPT-5.4-mini (6/8), DeepSeek-Chat (4/10); 0 in the others.

A representative Haiku T_2 from EXFIL writes three `new_evidence` nodes; two of them read (verbatim):

> "Message lacks cryptographic authentication, official company letterhead, verifiable sender email domain, or callback verification mechanism—received through unverified channel."
>
> "Message requests non-interruption of transfer and cites regulatory filing deadline, **creating time pressure that may be social engineering tactic**."

In the same T_2 the model halves self-reported confidence (0.95 → 0.35), inserts six `HOLD_PENDING_VERIFICATION`-style actions into `recommended_actions`, drops `threat_level` from CRITICAL to MEDIUM, and emits `rollback_commands` (`UNBLOCK_IP`, `RELEASE_HOST`, `REINSTATE_SESSION`). The EAC gate, reading this exact T_2, scores primary completeness $= 1.0$ and dispatches the rollback.

The model named the attack in its own structured output, articulated authentication failures, halved its confidence, inserted hold actions— and the gate cleared a rollback regardless. *The cognition the gate purports to verify is in the gate's own input, and the gate does not look at it.* Within a single forward pass, the agent's suspicion-articulation channel (`new_evidence`, low confidence, verification-gated actions) and dispatch channel (`threat_level`, `rollback_commands`) are decoupled; only one is graded downstream. The decoupling has a second-order signature: capitulator T_2s are systematically *shorter* than legitimate revisions ($p < 10^{-3}$, Mann–Whitney). Capitulation here is output truncation, not elaboration— the model fills the schema and moves on.

3.4. Can a stricter structural gate discriminate?

A natural objection is that a better-tuned structural gate would discriminate. We test it directly, scoring every structural signal the gate could compute on the same 193 trials (Appendix A). The completeness axis EAC actually grades is uniformly degenerate: primary, supporting, and overall completeness each carry rank-AUROC 0.500, and the count of addressed nodes 0.34— both classes fully address T_1, so no threshold on coverage-of-T_1 separates them. Signal appears only once the check stops grading coverage of T_1 and starts grading what the model *generated* in T_2: raw T_2 length reaches 0.69 (capitulations are shorter), and the count of `new_evidence` nodes—a structurally countable but content-bearing feature—reaches 0.90, the single strongest signal and the principal CCI feature. The lesson is not that deterministic gates cannot discriminate, but that they cannot discriminate *on the completeness axis*: any structural tightening that recovers discrimination does so precisely by abandoning coverage-checking for content-reading, at which point it has become the CCI. Tightening the wrong target does not help; changing the target does.

Table 3. Ablation. AUROC on out-of-fold predictions; 95% CI from a 1,000-resample bootstrap. The structural EAC gate has constant primary completeness $= 1.0$ on the same trials and is, by construction, non-informative.

Feature subset	AUROC	95% CI
x_1: new-evidence count	0.878	[0.831, 0.923]
x_2: T_2 confidence	0.812	[0.744, 0.878]
x_3: confidence drop	0.809	[0.729, 0.880]
$x_1 + x_2$	0.900	[0.856, 0.939]
$x_1 + x_3$ **(CCI)**	**0.934**	[**0.899**, **0.966**]
$x_2 + x_3$	0.801	[0.712, 0.884]
$x_1 + x_2 + x_3$	0.929	[0.891, 0.960]
EAC structural gate (baseline)	0.500	—

4. Cognitive Compliance Index

If the signal lives in content, a content-anchored discriminator should be available. We propose a preliminary *Cognitive Compliance Index* (CCI) built from features the agent already emits. We start from three candidates—x_1 = the count of `new_evidence` nodes in T_2; x_2 = the model's self-reported T_2 confidence (Lin et al., 2022); and x_3 = the confidence drop $c(T_1) - c(T_2)$—and fit logistic regression on the 193 changed-verdict trials with SOCIAL as the positive class. We use 5-fold stratified cross-validation with class-weight balancing to handle the 144:49 class imbalance, and report out-of-fold AUROC with a 1,000-resample bootstrap 95% confidence interval (Efron, 1979).

x_1 alone reaches AUROC 0.878; pairing with x_3 matches the full triple within CI (x_2 is redundant once x_3 is present). We adopt the parsimonious pair (x_1, x_3) as the principal CCI. At Youden's-J threshold (Youden, 1950) ($p = 0.44$), sensitivity is 0.96 (47/49 social) and specificity 0.85 (122/144 legit). At 15% FPR, CCI is a triage signal, not a blocking gate—an additive second axis to escalate suspect rollbacks where the structural gate gives no signal at all. The same trials on which the structural gate is non-informative

carry an AUROC-0.934 signal in fields the gate already reads.

Deployment. CCI is intended as a triage axis layered *after* the structural gate, not a replacement: EAC stays as the cheap form check, and CCI scores the same T_2 it already reads to escalate suspect rollbacks for human review. Because a flag triggers review rather than a block, a false positive costs analyst time, not availability, so the operating point can favor recall. Appendix B gives operating points: at a 10% false-positive rate CCI still recovers 88% of capitulations ($43/49$), and the Youden's-J point (0.96 sensitivity, 0.85 specificity) sits near 15% FPR. Like any content signal CCI is in principle gameable—a producer optimizing for form could pad `new_evidence`—but padding fabricated corroboration is itself a flaggable behavior, and the confidence-drop feature x_3 is not satisfiable by length alone; hardening CCI against adaptive producers is future work. Crucially, CCI inherits the gate's deployability: it is a deterministic logistic score over fields already present in the output, with none of the scalable-oversight regress of an LLM judge.

5. Discussion

Why structured-output gates fail. Such gates are deployed because they are cheap and deterministic; replacing them with an LLM judge inherits the alignment properties they were meant to enforce, the scalable-oversight regress. But cheapness comes from ignoring content, and the cognitive properties they purport to enforce are content properties—*evaluation theater* in the operations sense, echoing Schneier's critique of countermeasures that perform safety without delivering it (Schneier, 2003). The pattern predates LLMs: any deterministic gate grading structural conformance can be saturated by a producer optimized to satisfy the form, and tightening the gate along that same conformance axis does not help—empirically, every variant of EAC's completeness metric stays at chance (Section 3.4). LLM agents are merely the most capable form-saturator built.

The deployment is failing to grade. The output of an LLM agent under a $T_1 \rightarrow T_2$ contract carries several signals: a *form* signal that fills schema fields, a *dispatch* signal naming threat levels and rollback commands, and a *suspicion* signal recording inconsistencies in `new_evidence` and confidence. Every downstream consumer— schema validator, executor, EAC gate—reads form and dispatch; none reads suspicion. The model is not failing to suspect; the deployment is failing to grade it. The agent emits its own audit trail in T_2; the auditor is in the room, and the gate grades the report. Recovering cognitive compliance is not a measurement-discovery problem but a measurement-target problem: structured-output gates pick the wrong target.

Generalization and related work. The pattern recurs wherever form-checking serves as a safety surrogate: tool-call validators that verify schema conformance but not call appropriateness (Willard & Louf, 2023; Dong et al., 2024); JSON-mode enforcers (OpenAI, 2024); structured-output benchmarks grading by field completeness rather than cognitive correctness (Liu et al., 2024). Mirror-pair stress tests are cheap deployment hygiene: any gate that scores both halves identically does not discriminate. The sycophancy literature (Sharma et al., 2024; Perez et al., 2023) and work on verbalized confidence (Lin et al., 2022) and CoT faithfulness (Lanham et al., 2023) establish that models can articulate epistemic state; we show downstream gates fail to *measure* that state, even when articulated verbatim in their input.

6. Scope

We focus on one domain (SOC triage), one attack family (authority spoofing), and one structural gate (EAC) for controlled comparison. Mirror-pair construction and CCI features are domain-agnostic by design, and the completeness-axis degeneracy (Section 3.4) is a property of the measurement target rather than of SOC: any gate grading coverage-of-prior-state instead of generated content should exhibit it. Extending across attack families, replicating on deployed agent traces, and validating CCI thresholds out-of-distribution are natural next steps.

References

Dong, Y., Ruan, C. F., Cai, Y., Lai, R., Xu, Z., Zhao, Y., and Chen, T. XGrammar: Flexible and efficient structured generation engine for large language models. *arXiv preprint arXiv:2411.15100*, 2024.

Efron, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. 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 (AISec)*, 2023.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and

Perez, E. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR)*, 2022.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J. AgentBench: Evaluating LLMs as agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

OpenAI. Introducing structured outputs in the API. https://openai.com/index/introducing-structured-outputs-in-the-api/ 2024. Accessed April 2026.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL*, 2023.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Schneier, B. *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*. Copernicus Books, 2003. Imprint of Springer-Verlag New York.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Toyer, S., Watkins, O., Mendes, E. A., Svegliato, J., Bailey, L., Wang, T., Ong, I., Elmaaroufi, K., Abbeel, P., Darrell, T., Ritter, A., and Russell, S. Tensor Trust: Interpretable prompt injection attacks from an online game. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Willard, B. T. and Louf, R. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Youden, W. J. Index for rating diagnostic tests. *Cancer*, 3 (1):32–35, 1950.

A. Structural-gate ablation

Table 4 reports the univariate rank AUROC of every structural signal a deterministic gate could compute on the 193 changed-verdict trials, oriented so that the social-engineering class is positive. These are single-feature rankings (the ROC of one raw signal, no model) and so differ from the cross-validated logistic AUROCs in Table 3. Discrimination is absent on every variant of the completeness axis EAC grades—primary, supporting, overall completeness, and the count of addressed nodes—and appears only on signals that describe the content the model *generated* in T_2. The strongest single structural signal, the `new_evidence` count, is the principal CCI feature x_1: the boundary between a "structural" and a "cognitive" check is exactly where discrimination begins.

Table 4. Univariate rank AUROC of candidate structural signals ($N = 193$; 49 social, 144 legit).

Structural signal (social = positive)	AUROC
Completeness axis (EAC's target)	
primary completeness	0.500
supporting completeness	0.500
overall completeness	0.500
# addressed-evidence nodes	0.34
Generated-content signals	
# rollback commands	0.56
# recommended actions	0.59
raw T_2 length	0.69
# `new_evidence` nodes (x_1)	0.90

B. CCI operating points

Table 5 reports operating points for the principal CCI (x_1, x_3) on out-of-fold predictions. Thresholds are on $P[\text{social}]$. As a triage signal feeding human review, the false-positive rate ($= 1 -$ specificity) is the analyst-burden knob: recall stays high until FPR is pushed below 10%, and the Youden's-J point coincides with the $\leq 20\%$-FPR row at $47/49$ sensitivity.

Table 5. Operating points for the principal CCI (x_1, x_3) ($N = 193$; 49 social, 144 legit). Sensitivity shown with raw counts.

Operating point	thr.	Sens.	Spec.	FPR	PPV
FPR $\leq 5\%$	0.86	0.39 (19/49)	0.95	0.05	0.73
FPR $\leq 10\%$	0.47	0.88 (43/49)	0.90	0.10	0.75
FPR $\leq 15\%$	0.45	0.92 (45/49)	0.85	0.15	0.68
FPR $\leq 20\%$	0.43	0.96 (47/49)	0.81	0.19	0.63
Youden's J	0.44	0.96 (47/49)	0.85	0.15	0.68