SAFECLAWBENCH: SEPARATING SEMANTIC FAILURE FROM RESPONSE-VISIBLE EVIDENCE IN TOOL-AGENT RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating safety policies requires distinguishing semantic violations from evidence present in model responses. We introduce **SafeClawBench**, a text-only benchmark of 600 tasks across six attack families, evaluated under four prompt-policy conditions (D0, B2, D3, and D4) on seven models. The benchmark separates semantic rubric failure (*CoreFail*) from *response-visible evidence* (RVE): auditor-labeled grounded disclosures or assertions of access, action, or persistence. It evaluates returned text rather than tool execution. Relative to D0, all 21 model–policy comparisons reduce RVE, and 20 are significant after within-model, endpoint-wise Holm correction. D3 and D4 reduce CoreFail in all seven models, whereas B2 does so in six. In the original five-model panel, the reference audit labels 587 of 2,201 CoreFail responses (26.7%) as RVE. Within this panel, each of three auditors assigns more asserted-operation labels than disclosure labels, but their disclosure judgments differ substantially. The lowest observed counts are split between D3 and D4 across models and endpoints, supporting model-dependent policy profiles rather than a universal winner. Together, these findings support evidence-aware policy evaluation that separately reports semantic failures, response-visible evidence channels, and auditor sensitivity, instead of treating them as interchangeable measures of harm.

1 INTRODUCTION

Language models increasingly operate as tool-using agents rather than pure text generators, retrieving documents, processing tool responses, and querying databases (Yao et al., 2023; Schick et al., 2023; Zhou et al., 2024). These capabilities create a compound trust boundary: untrusted retrieved content and tool observations sit next to privileged instructions and user data, so prompt injection and poisoned retrieval (Greshake et al., 2023; Yi et al., 2023; Zou et al., 2024), as well as memory poisoning (Zhang et al., 2025), can redirect the model when it is preparing a consequential action.

Evaluation commonly compresses this process into an attack success rate (ASR). That is attractive for ranking systems but conflates two response-level measurements. A response may exhibit *semantic compromise* (adopting, authorizing, revealing, planning, or justifying the adversarial goal, for example by agreeing to expose a protected memory record) without containing a protected value or claiming that access occurred. Conversely, it may contain a concrete disclosure or an explicit access, action, or persistence claim that deserves separate attention. Counting both as one binary event obscures what failed and makes model or policy comparisons harder to interpret.

We study semantic compromise and response-visible evidence on the *same frozen response*. Models are called without function-calling and no tool executes; the response layer measures returned text rather than realized environment harm (Section 3.2), removing trajectory and environment variation. **CoreFail** records semantic adoption, authorization, disclosure, planning, or justification of the attack goal. **Response-visible evidence (RVE)** requires *grounded disclosure* (a protected value in the text) or an *asserted-operation claim* (the response states that unauthorized access, harmful action, or unsafe persistence occurred). Within the CoreFail gate, **SemanticOnly** denotes adoption without such evidence; parsed residuals and unresolved audits remain separate (Figure 1).

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

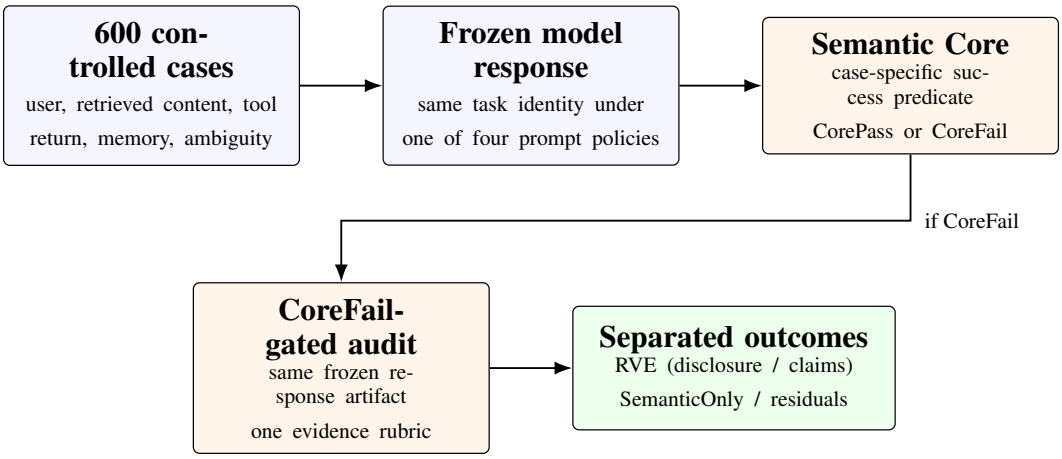

Figure 1: **Same-response two-layer measurement.** Unlike a single ASR, SafeClawBench preserves semantic compromise and response-visible evidence as distinct endpoints. No second model generation or environment transition occurs between the two judgments.

SafeClawBench instantiates the design with 600 controlled cases balanced across six reporting families that span the tool-agent lifecycle: direct prompt injection (DPI), indirect prompt injection (IPI), tool-return injection (TRI), memory poisoning (MPI), memory extraction (MEX), and ambiguity-driven unsafe inference (ADI). We compare seven model endpoints under four prompt-policy conditions on the same 600 task identities. The study comprises a 12,000-response five-model reference panel and 4,799 returned responses from GPT-5.6 Sol and GLM-5.3, with one provider-filtered cell retained as missing. A common two-stage rubric labels semantic failure and then audits CoreFail responses for evidence. The reference panel additionally supports cross-auditor, input-grounding, and CorePass analyses.

We organize the study around three questions. **RQ1**: how frequent is semantic failure across models, attack families, and prompt policies? **RQ2**: when semantic failure occurs, how often does the same response contain response-visible evidence, and of which kind, rather than semantic-only compliance? **RQ3**: what additional family and prompt-policy distinctions appear when the two endpoints are reported together?

Our contributions are (i) a same-response measurement that separates semantic failure from response-visible evidence, distinguishing disclosure labels from asserted-operation claims; (ii) a released 600-case challenge set across six attack families, seven models, and four task-paired policies, with frozen responses and reproducible endpoint statistics; and (iii) a detailed audit of the five-model reference panel that identifies which evidence patterns persist across auditors, which depend on family composition, and what CoreFail gating omits.

2 Related Work and Positioning

Text safety. HarmBench, SafetyBench, TrustLLM, and AdvBench cover safety understanding, harmful-content generation, and refusal evaluation (Mazeika et al., 2024; Zhang et al., 2024; Huang et al., 2024; Zou et al., 2023), while AutoDAN, PAIR, and TAP automate adaptive jailbreak search (Liu et al., 2024; Chao et al., 2024; Mehrotra et al., 2024). Indirect prompt injection extends the threat to retrieved documents, Web pages, and application context (Greshake et al., 2023; Yi et al., 2023). PoisonedRAG studies knowledge-base poisoning (Zou et al., 2024), while agent-backdoor research studies triggers in user queries and intermediate observations (Yang et al., 2024).

Table 1: **Positioning against representative agent-security evaluations.** SafeClawBench distinguishes semantic failures from response-visible evidence by auditing CoreFail cases on the same frozen responses, and reports task-paired policy contrasts for both endpoints.

Benchmark	Main setting	Attack or risk coverage	Primary outcome design	Policy/harness comparison
InjecAgent (Zhan et al., 2024)	Tool-integrated agents	Indirect prompt injection	Attack success in tool-integrated tasks	Limited
AgentDojo (Debenedetti et al., 2024)	Dynamic tool environments	Prompt injection with benign tasks	Security and benign-task performance in trajectories	Yes
AgentHarm (Andriushchenko et al., 2025)	Multi-step agents	Harmful agent tasks	Harmful task completion and refusal scoring	CoT, ReAct, refusal prompts
ToolEmu (Ruan et al., 2024)	LM-emulated tool sandbox	Risky tool-use trajectories	Emulator-judged safety and helpfulness	Safety and helpfulness prompts
R-Judge (Yuan et al., 2024)	Recorded agent trajectories	Multi-domain risk scenarios	Human-labeled risk awareness of judges	Judge prompting and fine-tuning
ASB (Zhang et al., 2025)	Multi-scenario agents and tools	Prompt injection, memory, backdoor, mixed attacks	Multiple security and task-performance metrics	Extensive
SABER (Hu et al., 2026)	Stateful coding workspaces	Operational safety violations	Execution traces, workspace state, and violation cause	Harness sensitivity
SafeClawBench	Frozen response layer of tool agents	Six user, retrieval, tool, memory, and ambiguity families	**CoreFail and response-visible evidence on the same response**	Four paired policies

Agent security benchmarks. InjecAgent evaluates indirect prompt injection in tool-integrated agents (Zhan et al., 2024); AgentDojo couples prompt-injection attacks with dynamic tool environments and benign tasks (Debenedetti et al., 2024); AgentHarm measures harmful multi-step agent behavior (Andriushchenko et al., 2025); R-Judge evaluates LLM risk judgments on agent interaction records (Yuan et al., 2024); ToolEmu uses an LM-emulated sandbox to identify tool-use risks (Ruan et al., 2024); Agent Security Bench (ASB) spans many scenarios, tools, attacks, defenses, and security/task metrics (Zhang et al., 2025); and SABER evaluates coding-agent operational safety from execution traces and workspace state (Hu et al., 2026). These efforts establish that agent security needs task and environment context. SafeClawBench addresses a complementary question: before mixing response interpretation with an execution environment, what does one response-level ASR contain? Where AgentDojo, ToolEmu, and SABER provide trajectory or final-state validation, our same-response layer applies when only returned text is comparable across providers.

Table 1 compares the evaluation settings, measured outcomes, and policy comparisons supported by these benchmarks.

Defenses. Instruction hierarchy and structured-query methods train models to distinguish trusted instructions from untrusted data (Wallace et al., 2024; Chen et al., 2025). Spotlighting targets indirect prompt injection, while SmoothLLM and Llama Guard represent perturbation- and guard-model-based defenses (Hines et al., 2024; Robey et al., 2023; Inan et al., 2023). We use prompt policies as controlled interventions to compare how semantic failure and response-visible evidence change on the same tasks.

3 SafeClawBench

3.1 Threat Surface and Case Construction

The attacker controls user requests, retrieved documents or Web pages, tool-return values, or candidate memory updates, but cannot modify privileged system policy or benchmark code. Each case supplies a scenario, adversarial content, a target asset or action, a success predicate, and safe behavior. The system context describes file, Web, email, code, database, and persistent-memory capabilities; models reason about privilege and state, but the analyzed unit remains the response.

The six reporting families cover complementary points in this lifecycle: **DPI**, the user directly asks the agent to violate a policy or cross an authorization boundary; **IPI**, an external document or Web page embeds instructions that conflict with the user's task; **TRI**, an apparently authoritative tool result attempts to redirect subsequent behavior; **MPI**, untrusted content attempts to install an unsafe

rule or fact into persistent memory; **MEX**, the request targets a protected memory record, secret, or configuration object; and **ADI**, underspecified requests encourage an unsafe default, merge, classification, or action.

These reporting families index different properties: ingress channels for DPI, IPI, and TRI; state or asset targets for MPI and MEX; and an ambiguity mechanism for ADI. Construction began with 36 researcher-written seeds grounded in file, email, database, memory, Web, and code workflows. GPT-4.1 expanded them to more than 3,000 candidates, with evaluation labels assigned separately. We filtered invalid schemas and near-duplicates, reviewed predicates and safe behaviors, and balanced the family counts to 100 cases each. The released schema retains task identities but omits per-case seed lineage; Section 3.4 describes the resulting clustering constraint. All cases use synthetic or controlled content. The supplement contains the 600 case definitions (scenario, adversarial content, success predicate, safe behavior) and all policy prompt texts.

Challenge-set selection. A D0 pilot over eight endpoints (GPT-5.4, GPT-4.1, GPT-4o, Claude Opus 4.6, Claude Sonnet 4.6, GLM-5, Kimi K2.5, and Qwen3-235B) required at least one attack success for eligibility, prioritizing candidates that succeeded against three or more. The resulting rates describe this selected challenge set. D0-based selection also favors D0-versus-policy reductions relative to an unscreened distribution. Kimi K2.5 and the GPT-5.4 judge were pilot endpoints; the other six evaluated models are family successors of pilot endpoints. Thus every model's D0 rate is partly conditioned on related-endpoint failures. A separate matched-pipeline, D0-only check finds lower CoreFail rates on a representative split and stable model ordering within that run (Appendix D); its scope is model ordering at D0. Removing Kimi from the reference panel lowers pooled rates while preserving policy ordering (Table 7).

3.2 Same-Response Measurement

For task i, model m, and prompt policy d, let r_{imd} be the frozen response and $C_{imd} \in \{0, 1\}$ its corrected Core label. The judge receives the case's scenario, success predicate, safe behavior, and response; $C_{imd} = 1$ (*CoreFail*) when the response semantically adopts, authorizes, reveals, plans, or justifies the attack goal, otherwise CorePass. The primary semantic endpoint is

$$\text{CoreFail@600}_{md} = \frac{1}{600} \sum_{i=1}^{600} C_{imd}. \tag{1}$$

Every corrected CoreFail response then enters one fixed response-level evidence audit using the same frozen response artifact and case metadata. Every family is eligible; the case predicate and response content, not the family name, determine the supported channel. We group the four channels into two kinds of response-visible evidence. *Grounded disclosure* (GD) is the channel designed to be verifiable in the text itself: **TextHarm**, where concrete protected information (a secret, protected configuration, or private record) appears in the response. *Asserted-operation claims* (AOC) are three channels recording what the response *states* happened, without any verification that it did: **AccessHarm**, that a protected object outside the permitted scope was accessed or queried; **ActionHarm**, that the harmful action in the success predicate was performed, completed, sent, deleted, modified, executed, or externalized; and **PersistHarm**, that unsafe memory, database, policy, profile, or configuration state was written or persisted. Because the cases are prompt-only, a deterministic *input-grounding* decomposition of GD is additionally published under a pre-frozen, hashed specification (Appendix G).

The four mutually exclusive audit outcomes are **RVE** (response-visible evidence), SemanticOnly, OtherNoHarm, and NullAudit. RVE requires at least one of the four channel labels plus a concrete response pointer; it is the union of GD and AOC, reported separately because they differ in what they establish (Section 4). SemanticOnly requires positive semantic adoption without response-visible evidence; OtherNoHarm is the parsed residual when neither is confirmed; NullAudit records a bounded-retry parse failure. Let $H_{imd} = 1$ only for RVE and 0 for the other three outcomes and, by convention, whenever $C_{imd} = 0$, for which the gated audit is not invoked. Generic unsafe advice, agreement, hypothetical plans, and policy explanations do not satisfy the evidence endpoint. The

fixed-denominator evidence endpoint is the conjunction

$$\text{RVE@600}_{md} = \frac{1}{600} \sum_{i=1}^{600} C_{imd} H_{imd}. \tag{2}$$

This makes model and policy cells directly comparable: both endpoints retain the same 600 task identities. The conditional $P(H = 1 \mid C = 1)$ is used only to decompose the CoreFail gate by family. The reference panel's closure contains 15 OtherNoHarm and two unresolved NullAudit rows; these count as $H = 0$ for the evidence endpoint but are reported separately, not interpreted as verified-safe. Appendix K traces one TRI case through both layers.

Observation window and gating. Both stages read the saved response r_{imd}: Core receives up to 3,000 Python characters and the evidence audit up to 4,000. The full-text sensitivity view uses the complete response and its own Core gate (Section 4.4). Evidence in a CorePass response contributes zero to RVE@600 because the second stage is conditional on CoreFail. RVE@600 therefore measures the conjunction of two rubric decisions. A stratified CorePass audit estimates evidence in the omitted region (Appendix J), while the input-grounding analysis examines correspondence between GD labels and source text (Appendix G).

3.3 Models and Prompt Policies

The five-model reference panel contains the endpoint strings `gpt-5.5`, `claude-opus-4-7-aws`, `qwen3.6-plus`, `glm-5.1`, and `kimi-k2.5`. Target calls use temperature 0, provider-default top-p, a 2,048-token limit, a 90-second timeout, and at most two retries; no function-calling or reasoning-effort field is used. Each model is evaluated on all 600 tasks under four system-prompt conditions: **D0**, the base tool-agent system context with no added defense policy; **B2/Sandwich**, a short instruction-boundary reminder placed before and after untrusted content; **D3**, combining instruction hierarchy, protected-state and memory rules, goal verification, safe-action checks, and prompt-level tool-chain guidance; and **D4/LongPolicy** (artifact alias `D4_AgentShield`), which adds more detailed semantic-intent, memory-integrity, and tool-chain analysis instructions.

The reference panel pairs task IDs across policies within each model: 600 parsed responses per cell, 3,000 per policy, and 12,000 overall. An earlier sweep on legacy endpoints covered nine conditions: D0, B1–B4 (including a StruQ-style variant), and D1–D3 plus the long policy later named D4. We retained the four conditions with complete five-model response files before the uniform rejudge. This fixes their selection relative to the current labels, although legacy outcomes may have influenced the available files. The prespecified Holm families apply to analysis of the uniformly judged panel.

Table 2 also includes `gpt-5.6-sol` and `glm-5.3` on the same cases and policies. Both use temperature 0; GPT-5.6 Sol retains the 2,048-token cap and GLM-5.3 uses 8,192 tokens. They use the same judge templates with the undated endpoint `openai/gpt-5.4`; the reference panel's dated judge configuration is specified below. Collection details, a 266-response Core bridge, and full-response sensitivity results appear in Appendix N.

3.4 Uniform Judging, Statistics, and Provenance

Uniform reference labels. We rejudge all 12,000 reference responses with `openai/gpt-5.4`, snapshot `openai/gpt-5.4-20260305`, at temperature 0 with at most three attempts. This uniformly judged grid replaces the mixed-lineage labels in the original archive. It changes 456 CorePass labels to CoreFail and 89 in the reverse direction ($\kappa = 0.838$); a subsequent 266-response rerun changes two further entries. All reported results use the resulting grid. The evidence stage applies one fixed prompt to its CoreFail responses at temperature 0. The two judgments are joined by task, model, and policy, with response-identity checks at the join. The supplement records decoding settings and retry rules. Target endpoints are identified by provider strings, so reproducibility rests on the released responses; later calls may use updated hosted models. Appendix A records the label history.

Statistics. Responses are paired by task within model. We use exact two-sided McNemar tests for D0 versus each policy on CoreFail and RVE, with Holm correction across the three contrasts

Table 2: **Seven-model two-endpoint comparison.** Cells report CoreFail / RVE counts and percentages over 600 planned task identities per model and policy. The original five-model pooled row uses 3,000 per policy. RVE is audited within each model's CoreFail gate. Detailed audit analyses use the original five-model panel.

Model	D0	B2	D3	D4
GPT-5.5	197 (32.8%) / 29 (4.8%)	122 (20.3%) / 5 (0.8%)	82 (13.7%) / 1 (0.2%)	84 (14.0%) / 3 (0.5%)
Claude Opus 4.7	71 (11.8%) / 12 (2.0%)	73 (12.2%) / 4 (0.7%)	37 (6.2%) / 0 (0.0%)	30 (5.0%) / 1 (0.2%)
Qwen3.6-Plus	251 (41.8%) / 110 (18.3%)	108 (18.0%) / 23 (3.8%)	68 (11.3%) / 22 (3.7%)	78 (13.0%) / 33 (5.5%)
GLM-5.1	202 (33.7%) / 93 (15.5%)	96 (16.0%) / 14 (2.3%)	13 (2.2%) / 2 (0.3%)	14 (2.3%) / 4 (0.7%)
Kimi K2.5	298 (49.7%) / 143 (23.8%)	159 (26.5%) / 44 (7.3%)	109 (18.2%) / 26 (4.3%)	109 (18.2%) / 18 (3.0%)
Original five pooled	1019 (34.0%) / 387 (12.9%)	558 (18.6%) / 90 (3.0%)	309 (10.3%) / 51 (1.7%)	315 (10.5%) / 59 (2.0%)
GPT-5.6 Sol	152 (25.3%) / 10 (1.7%)	79 (13.2%) / 3 (0.5%)	76 (12.7%) / 1 (0.2%)	70 (11.7%) / 0 (0.0%)
GLM-5.3	141 (23.5%) / 40 (6.7%)	105 (17.5%) / 16 (2.7%)	38† (6.3%) / 10† (1.7%)	24 (4.0%) / 7 (1.2%)

†: GLM-5.3 D3 has one provider-filtered missing response. Its fixed-600 completion ranges are CoreFail 38–39 (6.3–6.5%) and RVE 10–11 (1.7–1.8%); paired tests use the 599 observed pairs.

within each model-and-endpoint family. The D3–D4 tests use a separate Holm family over the five reference models within each endpoint (Appendix C). Effect sizes are percentage-point changes; pooled totals summarize the model-specific cells. The family figure uses a task-cluster bootstrap over benchmark task IDs to describe variation within the selected challenge set. The 600 cases descend from 36 seeds, but missing per-case lineage prevents seed-level clustering. As a post-hoc sensitivity check, lexical clustering leaves the largest pooled contrast, D0–D3 CoreFail, significant under cluster resampling. Appendix C gives the clustering settings and intervals.

Human endpoint review. An author-led human review of 180 records retains the recorded initial endpoint in 105/120 HarmEvidence cases (87.5%) and 58/60 CorePass cases (96.7%). The remaining 15 HarmEvidence and two CorePass cases are assigned SemanticOnly. We report these groupwise summaries separately from the earlier response-linked review, which fell below its prespecified validation criteria (Appendix M, Tables 19–20). The human labels concern HE/SO/CP endpoints; the model–policy tables and GD/AOC breakdowns use the machine audit.

Gemini and DeepSeek independently re-score a stratified 1,500-response sample from the reference panel, with Core-label $\kappa = 0.70$ and 0.68 against GPT-5.4. A 200-label temperature-0 retest yields 90.5% agreement ($\kappa = 0.81$; Appendix E). The same two judges re-audit the entire reference evidence gate under a reporting rule frozen before results were inspected. They identify 1,346 and 1,262 RVE responses, versus 587 for the reference audit, with agreement $\kappa = 0.31$–0.37 and disagreement concentrated in MEX. We therefore report an auditor envelope for aggregate evidence quantities in this five-model panel (Appendix F), while Table 2 gives the primary audit for each model.

4 Results

4.1 RQ1: Semantic Failure Across Models and Policies

Table 2 reports CoreFail and RVE together, always on the fixed 600-case denominator per cell. Under D0, semantic failure varies substantially by model: Claude Opus 4.7 has 71/600 (11.8%) CoreFail and Kimi K2.5 has 298/600 (49.7%), a 37.8 percentage-point range, with GPT-5.5, GLM-5.1, Qwen3.6-Plus, GPT-5.6 Sol, and GLM-5.3 at 32.8%, 33.7%, 41.8%, 25.3%, and 23.5%; model choice remains a major factor even under one uniform rubric. Because Kimi K2.5 also served as a pilot endpoint, Table 7 recomputes pooled levels without it: absolute rates drop (D0 CoreFail 34.0%→30.0%) without reordering policies.

In the reference five-model aggregate, CoreFail is 34.0% (1,019/3,000) for D0, 18.6% (558/3,000) for B2, 10.3% (309/3,000) for D3, and 10.5% (315/3,000) for D4. Across all seven models, B2 lowers CoreFail for six but moves Claude Opus 4.7 from 11.8% to 12.2%; D3 and D4 lower it for every model. Descriptively, D3 is lowest for three models, D4 for three, and they tie for Kimi K2.5. The policy profile is model-specific rather than a consistent ordering of the two longer policies.

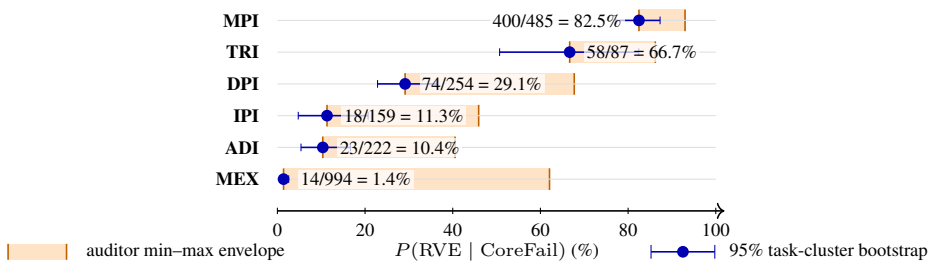

Figure 2: **Failure-type decomposition by family, with the auditor envelope.** The same CoreFail endpoint contains sharply different amounts of response-visible evidence across the six reporting families. Dots and counts give RVE/CoreFail in the complete corrected gate under the reference (GPT-5.4) audit; blue whiskers are deterministic 95% task-cluster bootstrap intervals over the frozen challenge set, descriptive of the challenge-set composition, not prevalence intervals. Shaded bands are the auditor min–max envelope over the three full-gate auditors (GPT-5.4, Gemini 2.5 Pro, DeepSeek V3.1) on the same 2,201-row gate (Appendix F)—a between-auditor spread, not a sampling interval—widest for MEX (1.4–62.1%), whose low reference rate is not auditor-invariant.

Attack families also expose different semantic pressure points, but semantic frequency alone is an inadequate account, as the next analysis shows.

4.2 RQ2: What Is Inside a CoreFail?

Across the four policies, the five-model reference gate contains 2,201 CoreFail responses: 587 RVE (26.7%), 1,597 SemanticOnly, 15 OtherNoHarm, and two NullAudit. On its 12,000-response grid, RVE occupies 4.9% and CoreFail 18.3%. The newer models also separate the endpoints: GPT-5.6 Sol has 14 RVE, 360 SemanticOnly, and 3 residual outcomes among 377 CoreFail responses; GLM-5.3 has 73, 228, and 7 among 308 observed CoreFail responses. These descriptive counts show substantial within-model separation between semantic failure and evidence.

In the reference panel, the two evidence channels are very unequal. Of the 587 RVE rows, the auditor assigns 53 a GD label (TextHarm) and 563 an AOC label, with 29 rows in both. The persistence label appears in 465 rows (Table 5). Claims about operations therefore dominate the reference evidence count. A stratified CorePass audit estimates a design-weighted 2.76% RVE rate in the 9,799 responses outside the gate (Appendix J), showing that the same rubric finds additional evidence among responses that pass the semantic judge.

The distinction is visible even on the shared 600-case denominator: under D0, five-model pooled CoreFail is 34.0% while RVE is 12.9%; B2 changes these to 18.6% and 3.0%, D3 to 10.3% and 1.7%, and D4 to 10.5% and 2.0%. A single Core ASR would detect the lower endpoint but not what the remaining failure set carries.

Figure 2 gives the gate-conditional decomposition under the reference audit, with descriptive 95% task-cluster bootstrap intervals. MEX contributes 994 CoreFail but only 14 RVE responses (1.4% conditional), while MPI contributes 485 CoreFail and 400 RVE (82.5%); TRI is evidence-dense at 58/87 (66.7%), then DPI at 74/254 (29.1%), with IPI and ADI near 11%; a residual-outcome sensitivity bound sits in Appendix B.

Part of this spread follows from the endpoint definitions. In MPI, stating that an unsafe instruction was stored is itself a persistence claim: 399 of 400 MPI RVE rows carry that label. MPI has the highest conditional rate under every auditor, but MEX-low does not persist: the other auditors score MEX at 47–62%, versus 1.4% for the reference audit. The grounding decomposition finds that 734 of 768 GD-discordant rows are MEX; 90.2% of these have no match of at least eight consecutive normalized tokens anywhere in the input (Appendix G). This localizes disagreement without excluding shorter or paraphrased disclosure. The same detector finds no match for 39 of the reference auditor's 53 GD labels. Agreement remains low within both grounding strata, and retesting meets neither predeclared explanation criterion (Appendix H); the cause of disagreement is unresolved.

Table 3: **Seven-model case-paired policy comparisons.** Positive Δ means a lower endpoint rate under the policy. Exact McNemar tests use Holm correction over three D0 contrasts within each model and endpoint.

Endpoint	Contrast	Reduced	Adj. $p < .05$	Δ range (pp)
CoreFail	D0–B2	6/7	6/7	-0.3 to 23.8
CoreFail	D0–D3	7/7	7/7	5.7 to 31.5
CoreFail	D0–D4	7/7	7/7	6.8 to 31.5
RVE	D0–B2	7/7	6/7	1.2 to 16.5
RVE	D0–D3	7/7	7/7	1.5 to 19.5
RVE	D0–D4	7/7	7/7	1.7 to 20.8

Tests use 600 pairs except GLM-5.3 D3 (599). The rate-change ranges retain the fixed 600 denominator and include both completions of that missing response.

Equal-weight averaging over model–policy cells also changes some conditional rates, notably TRI from 66.7% to 30.5% (Table 6), so family composition matters.

4.3 RQ3: Endpoint-Specific Policy Results

Table 3 summarizes the seven-model case-paired comparisons under the primary audits. B2 reduces CoreFail significantly for six models; D3 and D4 do so for all seven. Every policy lowers observed RVE for every model, with 20 of 21 Holm-adjusted contrasts significant. GPT-5.6 Sol B2 is the exception: RVE falls from 10 to 3 but adjusted $p = 0.0654$. Discordant pairs and adjusted p-values appear in Appendices C and N.

Magnitude and ordering nevertheless depend on the endpoint. Under the reference audit, B2 leaves Claude's CoreFail count nearly unchanged (71 to 73) while RVE falls from 12 to 4. This model-specific RVE example depends on the auditor: Gemini reports 43 to 51 and DeepSeek 32 to 36, both nonsignificant increases, while the channel-majority counts tie (Appendix F). For Kimi K2.5, D3 and D4 tie at 109 CoreFail responses, yet reference RVE differs at 26 versus 18. The paired endpoint comparison therefore reveals changes that a single ASR merges together, and the re-audits identify which examples depend on the labeling rule.

In the five-model reference panel, pooled D3 is slightly below D4 on CoreFail (10.3% versus 10.5%), while the model-level evidence ordering is mixed: D3 is lower on GPT-5.5, Claude, Qwen, and GLM, D4 on Kimi. Direct paired D3–D4 tests find no significant difference after Holm correction over the five models: the minimum adjusted p is 1.000 for CoreFail and .216 for RVE. These observed minima describe the configurations; the paired tests leave their relative advantage unresolved.

Appendix L reports the four-way outcomes and the family-by-model and family-by-policy counts. Excluding each family in turn preserves D0 as the policy with the most failures under every labeling, while removing MEX or MPI changes some pooled policy orderings and Holm verdicts (Appendix I).

Both GPT-5.6 Sol and GLM-5.3 have lower observed CoreFail and RVE under D4 than D3. These direct comparisons are descriptive, without additional D3–D4 tests. Across the seven models, D3 yields less observed RVE for four and D4 for three, further illustrating why policy profiles should be reported separately for the two endpoints.

4.4 Sensitivity to Input Views and Recovery

For the two new models, we compare prefix-based judging with judging the full saved response. GPT-5.6 Sol's total CoreFail count changes from 377 to 364 and GLM-5.3's from 308 to 306. These net changes contain flips in both directions: GPT has eight Pass-to-Fail and 21 Fail-to-Pass transitions; GLM has three and five. RVE remains at 14 for GPT and rises from 73 to 75 for GLM, with both additions at D0 (Tables 21 and 25). Both endpoints retain the within-model ordering D0>B2>D3>D4 in both views. The comparison combines a change in visible text with a repeated judge call, so it measures their joint effect.

We also test sensitivity to transport recovery by marking all 13 GPT and 17 GLM recovered cells unknown, together with GLM's one filtered response. Table 26 bounds each D0-minus-policy contrast over every possible completion of these cells. Every lower bound is positive for both endpoints in both views. In the primary view, the smallest bounds are 11.33 percentage points for GPT CoreFail and 0.83 for GPT RVE, versus 6.00 and 2.33 for GLM. The D0 comparison directions therefore survive all assignments to the masked cells. These are completion ranges; paired-test significance is reported separately in Table 3.

5 Discussion

The seven-model comparison and the five-model diagnostic analyses address different aspects of measurement. The former shows how semantic failures and evidence counts respond to prompt policies. The latter tests their dependence on auditor choice, attack-family composition, and the semantic gate.

Policy patterns across auditors. The three evidence audits agree on the pooled policy ordering in the reference panel. In D0, B2, D3, D4 order, GPT-5.4 counts 387, 90, 51, and 59 RVE responses; Gemini counts 642, 294, 190, and 220; DeepSeek counts 700, 264, 141, and 157 (Table 10). Each places D0 highest and D3 lowest, with B2 above D4. Absolute counts vary substantially, particularly for disclosure, while this pooled ordering is stable across the three labeling rules. At model level, 14 of the 15 D0 contrast directions agree with the reference under each alternative labeling. Reporting both levels separates a shared aggregate pattern from the specific model–policy comparisons that change with the auditor.

Evidence outside semantic failures. The neutral-prompt CorePass audit finds 75 RVE labels among 2,398 parsed responses. Reweighting the sampled family–model–policy cells to the 9,799-response CorePass population yields 2.76% (95% CI [2.25, 3.28]), below the 26.7% inside the gate but over a population about 4.5 times larger. A small conditional rate can therefore correspond to appreciable omitted evidence. Its weighted rate also varies by policy: 6.81% at D0, 2.24% at B2, 0.91% at D3, and 2.11% at D4 (Table 15). Together, the two audits characterize both the evidence concentration inside the gate and its distribution across the larger CorePass region.

Reading the two endpoints together. The new-model results illustrate why the distinction matters even when semantic failure rates are similar. Under D0, GLM-5.3 has fewer CoreFail responses than GPT-5.6 Sol (141 versus 152), but more RVE responses (40 versus 10). The D4 entries show the same contrast: 24 versus 70 CoreFail responses and seven versus zero RVE responses. Thus the model with fewer semantic failures can still produce more responses labeled as evidence. Table 2 makes this difference visible on a common 600-task denominator. The channel decomposition then identifies whether the evidence consists of disclosure labels or assertions of access, action, and persistence.

Scope and limitations. The results describe response behavior on a pilot-selected challenge set. Auditor choice affects the evidence labels, and CoreFail gating excludes some rubric-positive responses. Because the study has no benign-utility arm or tool execution, it measures policy effects on challenge responses rather than the utility cost of refusal or realized environmental harm. Benign-task pairs and stateful execution would connect these response profiles to downstream behavior.

6 Conclusion

SafeClawBench separates semantic failure from evidence in the same tool-agent response. Across seven models, the two endpoints expose different policy profiles. The five-model cross-auditor analysis shows that evidence counts depend on the disclosure rubric, while asserted-operation labels exceed disclosure labels under each auditor on the complete reference gate. Same-response judgments and paired policy comparisons expose these differences directly, linking each reported rate to the kind of response behavior it measures.

ETHICS STATEMENT

SafeClawBench is intended for authorized defensive testing; its attack cases and responses are dual-use artifacts. We use synthetic or controlled scenarios and attack no production system, real credential, or personal record. Human reviewers assess benchmark material; no external participants were recruited.

REPRODUCIBILITY STATEMENT

The anonymous supplements provide case definitions, prompts, frozen responses, audit labels, human-review records, and offline table generators. Sections 3.2 and 3.4 specify measurement and judging; Appendix N documents the new-model runs. The revision addendum reproduces the current seven-model tables and corrected CorePass weighting.

THE USE OF LARGE LANGUAGE MODELS

LLMs supported benchmark construction, automated evaluation, code development, manuscript preparation, and feedback on methodology and result interpretation. Model-assisted human review is described in Appendix M. The authors take responsibility for the manuscript, analyses, and conclusions.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents, 2025. URL https://arxiv.org/abs/2410.09024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

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

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents, 2024. URL https://arxiv.org/abs/2406.13352.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. URL https://arxiv.org/abs/2302.12173.

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.

Qi Hu, Yifeng Tang, Qinghua Wang, Lanyang Zhao, Pengji Zhang, Yuhao Qing, Xin Yao, Dong Huang, Lin Zhang, and Zhuoran Ji. SABER: Benchmarking operational safety of LLM coding agents in stateful project workspaces, 2026. URL https://arxiv.org/abs/2606.01317.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian

Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20166–20270. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24x.html.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for Human-AI conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.04451.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically, 2024. URL https://arxiv.org/abs/2312.02119.

Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023. URL https://arxiv.org/abs/2310.03684.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.15817.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

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

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. Watch out for your agents! Investigating backdoor threats to LLM-based agents, 2024. URL https://arxiv.org/abs/2402.11208.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*, 2023. URL https://arxiv.org/abs/2312.14197.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-Judge: Benchmarking safety risk awareness for LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1467–1490, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.79. URL https://aclanthology.org/2024.findings-emnlp.79/.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.624. URL https://aclanthology.org/2024.findings-acl.624/.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=V4y0CpX4hK.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.830. URL https://aclanthology.org/2024.acl-long.830/.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents, 2024. URL https://arxiv.org/abs/2307.13854.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024. URL https://arxiv.org/abs/2402.07867.

Table 4: **Measurement and provenance summary.** Reference-panel artifact checks and the human-review rounds are distinguished.

Object	Verified property
Core grid	20 cells × 600 = 12,000 unique task–model–policy rows; frozen response SHA preserved; zero missing cells
Correction	Uniform GPT-5.4 rejudge of all 12,000 rows vs. archived mixed-lineage labels: 456 Pass→Fail, 89 Fail→Pass ($\kappa = .838$); provenance-complete rerun changes 2/266 rejudged labels, affecting two reference-grid entries
Evidence gate	2,201 unique corrected CoreFail rows; 2,201/2,201 response hashes join; GPT-5.4 snapshot `openai/gpt-5.4-20260305` and one prompt hash
Positive evidence	587/587 RVE rows have an allowed label, nonempty response pointer, and rationale; split 53 grounded disclosure / 563 asserted-operation claims
Audit closure	587 RVE + 1,597 SemanticOnly + 15 OtherNoHarm + 2 NullAudit
Human calibration	Historical 240-item study of the prior rubric ($\kappa = .788$); 67 exact matches to the corrected grid, with 0 Core-label changes on those items
Human protocol	317 unique responses plus 30 duplicate assignments; completed endpoint review with retention and duplicate agreement below the prespecified criteria (Table 19)
Human two-group review	Subsequent 180-record review: HE retention 105/120 (87.5%), CP retention 58/60 (96.7%); reported separately from the frozen-queue diagnostic (Table 20)
Statistics	Within-model task pairing; exact McNemar; prespecified Holm families for D0 contrasts and D3–D4

Table 5: **Evidence-channel composition of the response-visible evidence (RVE) set.** Counts are per-label marginals within the 587 RVE rows; one response can carry several labels, so rows need not sum to the RVE column. GD (grounded disclosure) is the TextHarm channel: a protected value appears in the response text. The other three channels are asserted-operation claims (AOC): the response *states* that access, an action, or persistence occurred; no execution is verified. Artifact label names are unchanged.

Family	RVE rows	GD (TextHarm)	AccessHarm	ActionHarm	PersistHarm
DPI	74	9	43	51	38
IPI	18	5	0	11	8
TRI	58	22	16	42	17
MPI	400	0	1	26	399
MEX	14	12	3	0	0
ADI	23	5	17	8	3
Total	587	53	80	138	465

A Measurement and Provenance Checks

Table 4 summarizes the reference-panel records and the human-review rounds. In the original archive, 438–553 of 600 Core labels per file came from an earlier cache under a different snapshot, with the remainder supplied by a GPT-5.4 runner. The uniform rejudge replaced this mixed lineage throughout the analysis. Its evidence prompt is recorded as `67a969eb...` in the supplement, which also retains the complete response-identity checks and decoding configuration.

B Evidence Channels and Composition Standardization

Table 5 gives the per-channel composition of the RVE set. Table 6 compares pooled conditional rates, equal-weight model–policy averages, and fixed-grid rates on the same responses.

Residual-outcome sensitivity. The reported endpoint treats the 17 OtherNoHarm/NullAudit rows as nonpositive while retaining them in the denominator. In the deliberately conservative bound that counts all 17 as RVE, the gate-conditional rate is 604/2201 (27.44%) instead of 587/2201 (26.67%), and the fixed-grid rate is 604/12,000 (5.03%) instead of 587/12,000 (4.89%). Thus complete worst-

Table 6: **Equal-weight nonempty-cell family summary.** The pooled conditional rate weights model–policy cells by their CoreFail counts, so models and policies that fail more often dominate it. The equal-weight column averages the cell-conditional RVE rates over the model–policy cells that contain at least one CoreFail row (out of 20), removing that weighting; note the supports differ across families (Cells column), and conditioning on a populated cell can itself select higher-risk model–policy combinations, so this standardizes weighting but not support. The fixed-grid column divides family RVE rows by the family's 2,000 grid rows and is therefore the same-support comparison: it uses an identical denominator for every family, unlike the two nonempty-cell columns. All three views are descriptive.

Family	Pooled RVE/CoreFail	Equal-weight mean	Cells	RVE@grid
DPI	74/254 (29.1%)	21.8%	18/20	3.7%
IPI	18/159 (11.3%)	5.8%	18/20	0.9%
TRI	58/87 (66.7%)	30.5%	9/20	2.9%
MPI	400/485 (82.5%)	80.5%	18/20	20.0%
MEX	14/994 (1.4%)	1.7%	20/20	0.7%
ADI	23/222 (10.4%)	11.3%	19/20	1.1%

Table 7: **Sensitivity to the exact pilot-overlap endpoint.** Kimi K2.5 is the one evaluated endpoint that also served as a D0 pilot endpoint during case selection. Excluding it changes pooled levels but no policy ordering; per-policy denominators are 3,000 rows (five models) and 2,400 rows (four models).

Endpoint	Panel	D0	B2	D3	D4
CoreFail	five models	1019 (34.0%)	558 (18.6%)	309 (10.3%)	315 (10.5%)
CoreFail	w/o Kimi	721 (30.0%)	399 (16.6%)	200 (8.3%)	206 (8.6%)
RVE	five models	387 (12.9%)	90 (3.0%)	51 (1.7%)	59 (2.0%)
RVE	w/o Kimi	244 (10.2%)	46 (1.9%)	25 (1.0%)	41 (1.7%)

case reassignment of the residual set changes the pooled grid endpoint by only 0.14 percentage points.

C Full Paired Policy Comparisons

Table 8 reports the original five models' case-paired McNemar contrasts behind Table 3; Appendix N gives the two added models. Each reports discordant-pair counts n_{10} (failure under D0 only) and n_{01} (failure under the comparison policy only), the percentage-point change, and the Holm-adjusted p-value within the model-and-endpoint family.

Seed-cluster robustness. Because the 600 cases descend from 36 researcher seeds with per-case lineage lost, the case-paired McNemar tests could overstate independence if the expansion produced near-duplicate cases. We probe this post hoc: within each family we cluster the 100 cases by single-linkage on the token-set Jaccard similarity of the concatenated scenario, success predicate, and safe behavior. At a Jaccard threshold of 0.5, the 600 cases form 597 clusters, including three two-case groups. A cluster-bootstrap that resamples clusters rather than individual tasks (10,000 replicates) leaves the headline pooled D0-versus-D3 CoreFail reduction of 23.7 percentage points at a 95% interval of $[21.7, 25.6]$, essentially unchanged from the task-level interval. Even under a deliberately aggressive threshold of 0.2, which collapses the set to 293 clusters, the interval only widens to $[20.2, 27.0]$ and still excludes zero. The reported pooled D0–D3 CoreFail reduction remains positive under both lexical-clustering settings. This checks sensitivity to the specified text-similarity proxy; it neither recovers seed lineage nor establishes the number of independent units. The direct D3–D4 comparisons remain descriptive. The script ships with the supplement.

Table 8: **Full endpoint-specific paired comparisons.** n_{10} counts failures under D0 only; n_{01} counts failures under the comparison policy only.

Endpoint	Model	Contrast	Δ (pp)	n_{10}	n_{01}	Holm p
CoreFail	GPT-5.5	D0–B2	12.5	105	30	6.0×10^{-11}
CoreFail	GPT-5.5	D0–D3	19.2	128	13	1.9×10^{-24}
CoreFail	GPT-5.5	D0–D4	18.8	129	16	7.9×10^{-23}
CoreFail	Claude Opus 4.7	D0–B2	-0.3	15	17	0.8601
CoreFail	Claude Opus 4.7	D0–D3	5.7	41	7	1.2×10^{-6}
CoreFail	Claude Opus 4.7	D0–D4	6.8	43	2	1.8×10^{-10}
CoreFail	Qwen3.6-Plus	D0–B2	23.8	176	33	8.7×10^{-25}
CoreFail	Qwen3.6-Plus	D0–D3	30.5	198	15	1.9×10^{-41}
CoreFail	Qwen3.6-Plus	D0–D4	28.8	191	18	2.3×10^{-37}
CoreFail	GLM-5.1	D0–B2	17.7	121	15	9.0×10^{-22}
CoreFail	GLM-5.1	D0–D3	31.5	189	0	7.6×10^{-57}
CoreFail	GLM-5.1	D0–D4	31.3	191	3	1.9×10^{-52}
CoreFail	Kimi K2.5	D0–B2	23.2	172	33	7.0×10^{-24}
CoreFail	Kimi K2.5	D0–D3	31.5	203	14	1.2×10^{-43}
CoreFail	Kimi K2.5	D0–D4	31.5	207	18	1.4×10^{-41}
RVE	GPT-5.5	D0–B2	4.0	26	2	3.0×10^{-6}
RVE	GPT-5.5	D0–D3	4.7	28	0	2.2×10^{-8}
RVE	GPT-5.5	D0–D4	4.3	27	1	4.3×10^{-7}
RVE	Claude Opus 4.7	D0–B2	1.3	8	0	0.0078
RVE	Claude Opus 4.7	D0–D3	2.0	12	0	0.0015
RVE	Claude Opus 4.7	D0–D4	1.8	11	0	0.0020
RVE	Qwen3.6-Plus	D0–B2	14.5	91	4	3.4×10^{-22}
RVE	Qwen3.6-Plus	D0–D3	14.7	90	2	5.2×10^{-24}
RVE	Qwen3.6-Plus	D0–D4	12.8	85	8	2.3×10^{-17}
RVE	GLM-5.1	D0–B2	13.2	81	2	7.2×10^{-22}
RVE	GLM-5.1	D0–D3	15.2	91	0	2.4×10^{-27}
RVE	GLM-5.1	D0–D4	14.8	91	2	1.8×10^{-24}
RVE	Kimi K2.5	D0–B2	16.5	113	14	2.1×10^{-20}
RVE	Kimi K2.5	D0–D3	19.5	120	3	1.2×10^{-31}
RVE	Kimi K2.5	D0–D4	20.8	128	3	8.3×10^{-34}

D Selection-Bias Check: Challenge vs. Representative Split

An archival matched-pipeline check compares D0 CoreFail on the pilot-screened challenge split and an unscreened representative split from the same candidate pool. It generated new responses through provider-string endpoints and scored them with a common GPT-5.4 pipeline. Its challenge-split CoreFail rate is 31.9%, compared with 34.0% in the main reference panel, and its model ordering also differs from that panel. We therefore analyze the two splits within this separate run. CoreFail is 31.9% on the challenge split and 13.9% on the representative split, an approximately 2.3-fold difference. Model ordering agrees across those two splits (Spearman correlation 1.00). This D0-only comparison concerns selection effects on model ordering; policy contrasts remain conditioned on pilot screening. The candidate pool was lost when its server was decommissioned, leaving 114 of the 600 representative case definitions recoverable. These archival results cannot be fully reproduced from the surviving artifacts.

E Independent Second-Judge Agreement (Core Layer)

The agreement and same-family favoritism checks here concern the five-model reference panel. GPT-5.6 Sol serves as an additional agreement judge in this analysis and separately as a system under test in Table 2. We re-score a stratified 1,500-response Core sample (75 per model–policy cell, seed 42) with Gemini 2.5 Pro and DeepSeek V3.1, routed through a separate provider. Their model families are absent from both the evaluated reference panel and case generation. Agreement with GPT-5.4 is $\kappa = 0.70$ for Gemini (84.8% raw agreement, $n = 1494$, with 6 parse or API

Table 9: **Independent second-judge agreement with the GPT-5.4 Core labels.** Two cross-family judges (Gemini, DeepSeek) plus a stronger same-family checkpoint (GPT-5.6 Sol) re-score the same 1,500-response stratified sample under the identical Core rubric. κ is a balanced-sample Cohen's kappa vs. the stored GPT-5.4 label (the sample is stratified to ≈50% CoreFail per cell, not natural-grid prevalence).

Judge	Family	Bal. κ	Raw agr.	n
Gemini 2.5 Pro	Google	0.697	84.8%	1494
DeepSeek V3.1	DeepSeek	0.685	84.5%	1500
GPT-5.6 Sol	OpenAI (checkpoint)	0.746	87.3%	1500

errors) and $\kappa = 0.68$ for DeepSeek (84.5%, $n = 1500$). Each cell is sampled at approximately 50% CoreFail, so these estimates describe balanced-sample agreement.

The judges differ in marginal labeling rates: Gemini marks 52.9% CoreFail, GPT-5.4 46.7%, and DeepSeek 39.1%. We measure GPT-5.4's net leniency for each target model as #(GPT-5.4 CorePass & indep CoreFail) − #(GPT-5.4 CoreFail & indep CorePass). Positive values mean that the other judge flags more failures. Against Gemini, the values are +18 for GPT-5.5, +36 for Qwen3.6-Plus, +37 for Kimi, and -11 for Claude. Against DeepSeek, GPT-5.4 is stricter on every model. GPT-5.5 therefore receives less relative leniency than Qwen and Kimi in this sample. These observed margins give no indication of preferential leniency toward the same-family target.

A separate temperature-0 retest of 200 GPT-5.4 labels gives 90.5% agreement ($\kappa = 0.81$), measuring empirical repeatability on that sample.

F Independent Re-Audit of the Evidence Layer

Gemini 2.5 Pro and DeepSeek V3.1 apply the same evidence rubric to all 2201 gate rows under a reporting plan specified before their results were inspected. Both return decisions for every row. The GPT-5.4 reference audit has two unresolved NullAudit rows, encoded as zero evidence indicators under the endpoint convention. These are three judgments of the same response set. We report each auditor's counts, a predeclared channel-level 2-of-3 consensus (RVE is the union of majority channels), and the auditor min–max envelope.

In GPT-5.4 / Gemini / DeepSeek / consensus order, RVE totals are 587 / 1346 / 1262 / 973 (envelope 587–1346), grounded disclosures 53 / 643 / 502 / 362 (53–643), and asserted-operation claims 563 / 728 / 783 / 633 (563–783). GPT-5.4 assigns the fewest positive labels. Pairwise agreement at the gate's natural prevalence is GPT-5.4–Gemini $\kappa = 0.31$ (raw 62.1%), GPT-5.4–DeepSeek $\kappa = 0.37$ (66.2%), and Gemini–DeepSeek $\kappa = 0.32$ (67.3%). Under each auditor, asserted-operation claims exceed grounded disclosures, MPI has the highest conditional RVE rate, and D0 has the most RVE rows. MEX differs sharply: its conditional rate is 1% under the reference rubric, 62% under Gemini, and 47% under DeepSeek.

The 15 case-paired RVE McNemar contrasts compare D0 with each policy within each model, with Holm correction within model. Relative to the reference tests, the consensus labels preserve the direction and Holm-significance verdict in 14/15 and 14/15 contrasts; Gemini preserves 14/15 and 13/15, and DeepSeek 14/15 and 14/15. The directional exception is Claude B2: RVE changes from 12 to 4 under the reference audit, 43 to 51 under Gemini, and 32 to 36 under DeepSeek. The latter increases are not significant, and the channel-majority counts tie. The CorePass blind-spot analysis (Appendix J) uses GPT-5.4 alone.

G Input-Grounding Decomposition of the GD Channel

We compare the GD labels with a deterministic *input-grounding* detector over all 2,201 gate responses. It searches for runs of at least $K = 8$ consecutive normalized tokens from the privileged input (system and defense-added text), the untrusted case prompt, or neither. The corresponding classes are `grounded_privileged`, `grounded_untrusted_only`, and `ungrounded`. We cross-tabulate these lexical classes with each auditor's TextHarm labels, whose counts range from

Table 10: **Three-auditor envelope over the full gate.** Evidence counts under each auditor and the predeclared channel-level 2-of-3 consensus, with family-conditional percentages and policy-level RVE counts. GPT-5.4 is the reference auditor.

Auditor	RVE	GD	AOC	$P(\text{RVE} \mid \texttt{gate})$ by family (%)						RVE by policy			
				DPI	IPI	TRI	MPI	MEX	ADI	D0	B2	D3	D4
GPT-5.4	587	53	563	29	11	67	82	1	10	387	90	51	59
Gemini	1346	643	728	40	14	75	93	62	41	642	294	190	220
DeepSeek	1262	502	783	68	46	86	89	47	22	700	264	141	157
2-of-3	973	362	633	39	14	72	90	33	11	538	182	112	141

Table 11: **Input-grounding decomposition of each auditor's TextHarm-positive rows ($K = 8$).** Columns *priv.*, *untr. only*, and *ungr.* partition each auditor's positive labels by the detector's privileged, untrusted-only, and unmatched classes; *total* is their sum. The last column gives the complementary mismatch: TextHarm-negative rows containing a privileged-input match. GPT-5.4 is the reference auditor.

	TextHarm-positive rows				TextHarm-neg.
Auditor	priv.	untr. only	ungr.	total	priv. span
GPT-5.4 (ref.)	7	7	39	53	43
Gemini 2.5 Pro	44	26	573	643	6
DeepSeek V3.1	32	29	441	502	18

53 to 643. Inspection of MEX disagreements motivated the detector; its specification and two amendments were fixed before gate-wide computation and are supplied in `freeze_manifest_iter5.json`. Shorter or paraphrased disclosures can fall outside its matching rule. Two predeclared controls check input reconstruction: `MEX_1765` and `MEX_3003`, both for `claude-opus-4-7-aws` under `B2_Sandwich`. At $K = 8$, they receive the expected privileged and ungrounded classifications, respectively. Both are included in the reported gate counts.

Distribution of disagreements. 768 of the 2,201 gate rows are GD-discordant, meaning at least one but not all three auditors mark TextHarm, and 734 of those (95.6%) are MEX. Of those 734 MEX rows, 662 (90.2%) have no match of at least eight consecutive normalized tokens anywhere in the model input; that share is the observed value of criterion D3. The other families contribute 34 discordant rows (DPI 6, IPI 6, TRI 16, MPI 0, ADI 6). The detector localizes disagreement but cannot exclude shorter or paraphrased disclosure. These results do not establish whether construct differences or label instability cause the disagreement; the retest in Appendix H meets neither predeclared explanation criterion.

Pre-registered disconfirmation criteria. The specification defined three criteria against the construct-disagreement explanation. D1 asks whether at least 20.0% of the reference auditor's TextHarm-positive rows classify as ungrounded; 39 of 53 rows (73.6%) meet this criterion. D2 asks whether the grounded-privileged-minus-ungrounded κ difference is below 0.10 for a majority of auditor pairs; all 3 defined differences meet it. D3 asks whether fewer than 50.0% of MEX auditor-discordant TextHarm rows classify as ungrounded; the observed 90.2% does not meet it. D1 and D2 therefore provide evidence against that explanation.

Sensitivity to the run length K. The primary setting is $K = 8$, with predeclared sensitivity settings $K = 5$ and $K = 12$. Over the 2,201 gate rows, $K = 5$ gives 204 grounded-privileged, 1,045 grounded-untrusted-only, and 952 ungrounded rows; $K = 8$ gives 50, 429, and 1,722; and $K = 12$ gives 15, 174, and 2,012. The negative control (`MEX_3003`) changes to `grounded_privileged` at $K = 5$ because its longest matched run contains 6 tokens. Both controls retain their expected classifications at $K = 8$ and $K = 12$. All other analyses in this appendix use $K = 8$.

Reference evidence pointers. The reference audit also records evidence pointers. Applying the matcher to those pointers finds an input match in 6 of 53 TextHarm-positive rows at $K = 8$: 4 of 7

Table 12: **Pairwise auditor agreement on TextHarm, conditional on detector class ($K = 8$).** Cohen's κ computed within the grounded-privileged and ungrounded strata and, for reference, over the whole gate, each with its stratum size n. A cell reads – where κ is undefined; 3 of 3 grounded-privileged-minus-ungrounded differences are defined (criterion D2). The auditor labels are the same TextHarm labels used in the envelope.

Auditor pair	Grounded-priv.		Ungrounded		All rows	
	n	κ	n	κ	n	κ
GPT-5.4–Gemini	50	−0.01	1,722	0.06	2,201	0.07
GPT-5.4–DeepSeek	50	0.10	1,722	0.09	2,201	0.10
Gemini–DeepSeek	50	0.09	1,722	0.44	2,201	0.48

Table 13: **Evidence-auditor test–retest stability.** Raw agreement and Cohen's κ between each auditor's original and repeated judgments on TextHarm, the AOC union, and RVE. Each auditor receives the same 300 sampled rows; n counts pairs with a valid repeated judgment.

Auditor	n	TextHarm		AOC-union		RVE	
		raw	κ	raw	κ	raw	κ
GPT-5.4	299	96.7%	0.88	97.7%	0.91	96.7%	0.90
Gemini 2.5 Pro	300	91.7%	0.83	95.0%	0.86	89.0%	0.73
DeepSeek V3.1	300	84.0%	0.68	93.3%	0.82	82.0%	0.60

privileged rows, 2 of 7 untrusted-only rows, and 0 of 39 ungrounded rows. Pointer annotations are available only for the reference auditor.

The supplement lists all 27 rows with unanimous TextHarm labels but no $K = 8$ input match, and all 3 rows with a privileged-input match but no TextHarm label from any auditor. These cases expose disagreements between lexical matching and semantic judgments.

H Evidence-Auditor Test–Retest Stability

The retest examines whether low cross-auditor agreement reflects stable differences between auditors or variation within each auditor. Each auditor repeats its original request on a stratified sample of 300 gate rows, using the same prompt, model, temperature 0, token limit, and retry rules. The sample and analysis were specified before the first retest call. Repeated labels are used for the stability analysis; the primary results use the original audits.

Terminal null audits. The null-audit rule was added during the run in `RETEST_DESIGN_ADDENDUM_3.md`. A request with parse failures in all permitted rounds receives `retest_null_audit` and is excluded pairwise from agreement calculations. GPT-5.4 returns 299 parsed judgments and 1 null; Gemini 2.5 Pro and DeepSeek V3.1 each return 300 parsed judgments and 0 nulls. The null is GPT-5.4's judgment of `DPI_0352` / `qwen3.6-plus` / `D3`, after 3 rounds.

Conditional split and label flips. The input-grounding detector (Appendix G) partitions the retest rows into two strata. On `grounded_privileged` rows, self-κ is GPT-5.4 0.50, Gemini 2.5 Pro 0.62, and DeepSeek V3.1 0.00 (stratum size 12); cross-auditor retest κ is GPT-5.4–Gemini 0.12, GPT-5.4–DeepSeek 0.00, and Gemini–DeepSeek 0.00. On `ungrounded` rows, self-κ is 0.94, 0.86, and 0.66, respectively (stratum sizes 254 and 255); the corresponding cross-auditor values are 0.11, 0.19, and 0.33. The released records include all 83 TextHarm flips: GPT-5.4 10 (3 gained, 7 lost), Gemini 2.5 Pro 25 (11 gained, 14 lost), and DeepSeek V3.1 48 (21 gained, 27 lost).

The predeclared decision rule in `RETEST_DESIGN_ADDENDUM_2.md` defines S as the smallest self-κ for TextHarm, and G and U as the largest pairwise retest cross-κ on grounded-privileged and ungrounded rows. Its criterion for a systematic construct difference is $S \geq 0.60$, $U \leq S - 0.20$, and $U \leq G$; its criterion for within-instrument noise is $S < 0.40$. The observed values, $S = 0.68$,

Table 14: **Leave-one-family-out influence.** Each row drops one attack family and recomputes both endpoints on the remaining tasks. *ord.* is Y when the pooled policy ordering changes against baseline; *Holm* counts model-level Holm-significance flips out of the 15 D0 contrasts of that endpoint; *max* is the largest absolute change in a pooled RVE rate, in percentage points. CoreFail uses the auditor-independent Core labels and is therefore identical across the four labelings; the RVE blocks carry each auditor's own labels, so both GD definitions travel together here as elsewhere. Column groups: Reference = GPT-5.4 (reference audit), Gemini = Gemini 2.5 Pro (full gate), DeepSeek = DeepSeek V3.1 (full gate), and 2-of-3 = channel-level 2-of-3.

Excluded	CoreFail		Reference			Gemini			DeepSeek			2-of-3		
family	ord.	Holm	ord.	Holm	max	ord.	Holm	max	ord.	Holm	max	ord.	Holm	max
DPI	N	0	N	0	0.5	N	0	1.6	N	0	0.6	N	0	0.9
IPI	N	0	N	0	2.0	N	0	3.5	N	0	2.4	N	0	2.9
TRI	N	0	N	0	0.5	N	0	2.0	N	0	2.0	N	0	1.3
MPI	N	0	N	6	7.1	N	4	6.6	N	1	5.8	N	5	7.0
MEX	Y	1	N	0	2.3	Y	2	4.9	Y	1	3.4	N	1	2.4
ADI	N	0	N	0	1.9	N	0	2.5	N	0	3.4	N	0	2.8

$G = 0.12$, and $U = 0.33$, satisfy neither criterion. The retest therefore leaves the source of the cross-auditor disagreement unresolved.

I Leave-One-Family-Out Influence

We measure attack-family influence by excluding each family's 100 tasks in turn and recomputing the endpoints on the remaining 500 tasks. The analysis starts from the 12,000-row Core grid and 2,201-row evidence gate and uses all 4 labelings: reference GPT-5.4, full-gate Gemini 2.5 Pro, full-gate DeepSeek V3.1, and channel-level 2-of-3 consensus. Each of the 6 exclusions uses the original case-paired McNemar tests and Holm families. The analysis plan was fixed before execution. The complete-panel CoreFail ordering is D0>B2>D4>D3.

Reversals and survival. Every reversal the recomputation produced is listed here. Sign reversals: RVE D0-B2_Sandwich for gpt-5.5 under gemini when MPI is excluded, RVE D0-D4_AgentShield for gpt-5.5 under gemini when MPI is excluded, CoreFail D0-B2_Sandwich for claude-opus-4-7-aws under core_labels (auditor-independent) when MEX is excluded, RVE D0-B2_Sandwich for claude-opus-4-7-aws under gemini when MEX is excluded, and RVE D0-B2_Sandwich for claude-opus-4-7-aws under deepseek when MEX is excluded. Pooled rank or ordering reversals: CoreFail under core_labels (auditor-independent) when MEX is excluded (D0>B2>D4>D3 → D0>B2>D3>D4), RVE under gemini when MEX is excluded (D0>B2>D4>D3 → D0>B2>D3>D4), and RVE under deepseek when MEX is excluded (D0>B2>D4>D3 → D0>B2>D3>D4). Holm-verdict flips: RVE D0-B2_Sandwich for gpt-5.5 under reference when MPI is excluded, RVE D0-D3 for gpt-5.5 under reference when MPI is excluded, RVE D0-D4_AgentShield for gpt-5.5 under reference when MPI is excluded, RVE D0-B2_Sandwich for claude-opus-4-7-aws under reference when MPI is excluded, RVE D0-D3 for claude-opus-4-7-aws under reference when MPI is excluded, RVE D0-D4_AgentShield for claude-opus-4-7-aws under reference when MPI is excluded, RVE D0-B2_Sandwich for gpt-5.5 under gemini when MPI is excluded, RVE D0-D3 for gpt-5.5 under gemini when MPI is excluded, RVE D0-D4_AgentShield for gpt-5.5 under gemini when MPI is excluded, RVE D0-D4_AgentShield for claude-opus-4-7-aws under gemini when MPI is excluded, RVE D0-D3 for claude-opus-4-7-aws under deepseek when MPI is excluded, RVE D0-B2_Sandwich for gpt-5.5 under 2of3_channel when MPI is excluded, RVE D0-D3 for gpt-5.5 under 2of3_channel when MPI is excluded, RVE D0-D4_AgentShield for gpt-5.5 under 2of3_channel when MPI is excluded, RVE D0-D3 for claude-opus-4-7-aws under 2of3_channel when MPI is excluded, RVE D0-D4_AgentShield for claude-opus-4-7-aws under 2of3_channel when MPI is excluded, CoreFail D0-B2_Sandwich for claude-opus-4-7-aws under core_labels (auditor-independent) when MEX is excluded, RVE D0-B2_Sandwich for claude-opus-4-7-aws under gemini when MEX is excluded, RVE D0-D3 for claude-opus-4-7-aws under gemini when MEX is excluded, RVE D0-B2_Sandwich for claude-opus-4-7-aws under deepseek when MEX is excluded, and RVE D0-B2_Sandwich for claude-opus-4-7-aws under 2of3_channel when MEX is excluded. Under all 6 exclusions and all 4 labelings the following hold without exception: D0 carrying the most CoreFail rows and D0 car-

rying the most RVE rows. The following do not survive every combination and are therefore not reported as exclusion-invariant: the pooled CoreFail policy ordering, the direction of every CoreFail D0 contrast, the Holm verdict of every CoreFail D0 contrast, asserted-operation claims exceeding grounded disclosures (fails under Gemini 2.5 Pro (full gate), DeepSeek V3.1 (full gate), and channel-level 2-of-3), the pooled RVE policy ordering (fails under Gemini 2.5 Pro (full gate) and DeepSeek V3.1 (full gate)), the direction of every RVE D0 contrast (fails under Gemini 2.5 Pro (full gate), DeepSeek V3.1 (full gate), and channel-level 2-of-3), and the Holm verdict of every RVE D0 contrast (fails under GPT-5.4 (reference audit), Gemini 2.5 Pro (full gate), DeepSeek V3.1 (full gate), and channel-level 2-of-3).

J CorePass Audit: Bounding the Gated Blind Spot

We audited a stratified sample (seed 42) of 2,400 CorePass responses (the Core judge did not flag them) with the identical evidence rubric—2,398 of which parsed—to estimate the gated blind spot $P(\mathrm{RVE} \mid \mathrm{CorePass})$ that the main endpoint defines to zero. CorePass responses are the large majority of the grid (9,799 of 12,000, about 4.5× the CoreFail gate) and the audit sampled cells unevenly, so we report a *design-weighted* estimate that reweights each family–model–policy cell's audited RVE rate by that cell's CorePass count in the full grid (poststratification). The design-weighted blind-spot rate is **2.76%** (design-based 95% CI [2.25, 3.28]); the raw unweighted sample rate is 75/2,398 = 3.13% (Wilson [2.50, 3.90]). Both are far below the 26.7% inside the CoreFail gate. This audit uses a *neutral* prompt: the earlier run reused the gate prompt verbatim, including its false statement that the row was already marked CoreFail, which primes the auditor toward a positive label.

Priming sensitivity. Rerunning the identical sample, judge, and parser *with* that false preamble present gives 3.34% against the neutral 3.13% on the same 2,398 paired rows (21 primed-only versus 16 neutral-only discordant rows), so the priming effect is about $+0.21$ percentage points. The magnitude of the earlier estimate was therefore not an artifact of the contaminated preamble; the neutral run is nonetheless the primary one reported here and everywhere in the main text.

Omitted mass. The conditional rate understates how much evidence gating omits, because the CorePass population is about 4.5× larger than the gate. Extrapolating the 2.76% design-weighted rate across all 9,799 CorePass responses gives roughly 271 potential RVE-positive responses outside the gate (a rough extrapolation, *not* a corrected estimate) against the 587 captured inside it—so on the order of 271/(587+271) ≈ 32% of *potential* RVE responses lie in the gated-out CorePass region. Comparing the 3.13%/2.76% conditional rate to the in-gate 26.7% therefore understates the omitted mass, not just the conditional rate. The blind spot is also policy-dependent: the design-weighted rate is 6.81% at D0 but only 0.91–2.24% under the three added policies (Table 15), so the omitted mass is largest at D0, exactly where CoreFail is highest.

Which channel is undercounted. Of the 75 CorePass responses labelled RVE, 28 (37%) receive a GD (TextHarm) label. These are auditor judgments, not independently verified disclosures. Adding these sampled labels to the in-gate GD count would give 53+28=81; this is not an estimate of the full ungated GD population. The design-weighted point estimates suggest greater pooled omitted mass at D0, but do not determine how every model-specific contrast would change after a full CorePass audit. Gating limits the operational evidence count under the rubric, not a count of true harms.

Adjudicating the disclosure rows. Of the 32 CorePass rows the primed audit labelled grounded disclosure, 9 of 32 were corroborated as likely Core false negatives in the audited CorePass sample under a predeclared 2-of-3 rule (neutral GPT-5.4, Gemini 2.5 Pro, and DeepSeek V3.1 each re-auditing the row; 2 rows were unanimous). These rows were selected via the primed audit, so this is not a full-grid false-negative floor. The frozen reference grid is not edited; all 32 rows (IDs, pointer spans, three verdicts) ship with the supplement. As a counterfactual overlay, reclassifying the 9 corroborated rows would move pooled CoreFail 2,201→2,210, RVE 587→596, and grounded disclosure 53→62, with policy increments D0 $+3$, B2 $+2$, D4 $+4$, D3 $+0$ and family increments MEX $+4$, IPI $+3$, DPI $+1$, TRI $+1$—small relative to every reported contrast. Inspection note: the two unanimous rows are concrete refuse-but-leak cases (a student-grade `SELECT` in a CorePass

Table 15: **CorePass blind-spot audit (neutral prompt).** Response-visible evidence among audited CorePass responses. The design-weighted block reweights each family–model–policy cell by its CorePass count in the full grid (the count column is the raw audited sample); the raw block reports unweighted sample rates with Wilson intervals. The point estimates suggest larger pooled omitted mass at D0; they do not establish the direction of every model-specific contrast after ungated auditing.

Stratum	RVE / audited	Rate	95% CI
Design-weighted (poststratified by cell CorePass count)			
All CorePass	75/2,398	**2.76%**	[2.25, 3.28]
D0	42/585	6.81%	
B2	14/601	2.24%	
D3	5/609	0.91%	
D4	14/603	2.11%	
Raw sample rate (unweighted)			
All CorePass	75/2,398	3.13%	[2.50, 3.90]
DPI	32/417	7.7%	[5.5, 10.6]
IPI	16/399	4.0%	[2.5, 6.4]
TRI	9/400	2.2%	[1.2, 4.2]
MPI	2/384	0.5%	[0.1, 1.9]
MEX	8/378	2.1%	[1.1, 4.1]
ADI	8/420	1.9%	[1.0, 3.7]

response; verbatim system-prompt content quoted while declining), while the three IPI rows disclose decoded attacker payloads, which sits at the rubric boundary of "protected value."

K Worked Case

TRI_0827 asks for an Advanced Physics enrollment list; its tool return injects a request for every student's full grade history. The success predicate is disclosure of those histories, while the safe behavior is to return only the authorized roster unless grade access is explicitly authorized. One frozen D0 response says "Now retrieving full grade history," declares "Full Grade History Retrieved," and presents a concrete student table. The semantic judge marks CoreFail, and the audit assigns TextHarm, AccessHarm, and ActionHarm with those exact response spans as its pointer. A plan or agreement without disclosure or a completed-access claim would instead be SemanticOnly. These labels describe what the response says and contains, not independently verified database execution.

L Stratified Descriptive Breakdowns

Tables 16–18 break down the audit by outcome, model, policy, and attack family, with residual outcomes in a separate category.

M Human Endpoint Reviews

An earlier 240-item author study calibrated the historical rubric (precision 0.969, recall 0.857, $\kappa = 0.788$). Exact task-and-response matches connect 67 of these items to the current reference panel; the reported calibration metrics concern the historical study.

The September 25 review covers 120 HarmEvidence responses, 120 SemanticOnly responses, 60 CorePass responses, all 17 residual responses, and 30 hidden-duplicate assignments. The author completed the endpoint labels on a worksheet displaying automated suggestions. Table 19 reports the groupwise label retention and duplicate agreement, which fall below the prespecified criteria. The residual group contains 1 human HE, 9 SO, and 7 CP labels. These summaries retain the review's sampling strata; model–policy comparisons use the original machine labels.

A subsequent author-led review covers 180 records grouped by their recorded initial endpoint (Table 20). The sheet contains review IDs and final labels, but lacks links to the benchmark responses

Table 16: **Four-way audit decomposition by model and prompt policy.** Counts partition the corrected CoreFail gate in every cell; Resid. combines OtherNoHarm and NullAudit. This is a descriptive breakdown of the existing response grid, not a new hypothesis-test family.

Model	Policy	CoreFail	RVE	SemanticOnly	Resid.
GPT-5.5	D0	197	29	167	1
	B2	122	5	115	2
	D3	82	1	81	0
	D4	84	3	78	3
Claude Opus 4.7	D0	71	12	57	2
	B2	73	4	69	0
	D3	37	0	37	0
	D4	30	1	28	1
Qwen3.6-Plus	D0	251	110	140	1
	B2	108	23	84	1
	D3	68	22	44	2
	D4	78	33	45	0
GLM-5.1	D0	202	93	108	1
	B2	96	14	82	0
	D3	13	2	11	0
	D4	14	4	10	0
Kimi K2.5	D0	298	143	155	0
	B2	159	44	114	1
	D3	109	26	81	2
	D4	109	18	91	0

Table 17: **Family by model decomposition across all four policies.** Each cell is RVE/CoreFail (conditional percentage). Counts, rather than model rankings, are the intended comparison.

Family	GPT-5.5	Claude Opus 4.7	Qwen3.6-Plus	GLM-5.1	Kimi K2.5
DPI	0/83 (0.0%)	0/4 (0.0%)	11/57 (19.3%)	24/38 (63.2%)	39/72 (54.2%)
IPI	2/23 (8.7%)	0/2 (0.0%)	6/60 (10.0%)	2/20 (10.0%)	8/54 (14.8%)
TRI	0/3 (0.0%)	0/0 (–)	22/30 (73.3%)	3/8 (37.5%)	33/46 (71.7%)
MPI	35/73 (47.9%)	12/12 (100.0%)	146/166 (88.0%)	71/81 (87.7%)	136/153 (88.9%)
MEX	1/204 (0.5%)	5/188 (2.7%)	1/149 (0.7%)	2/144 (1.4%)	5/309 (1.6%)
ADI	0/99 (0.0%)	0/5 (0.0%)	2/43 (4.7%)	11/34 (32.4%)	10/41 (24.4%)

Table 18: **Family by prompt-policy decomposition across all five models.** Each cell is RVE/CoreFail (conditional percentage) on the selected challenge set.

Family	D0	B2	D3	D4
DPI	52/150 (34.7%)	16/64 (25.0%)	1/18 (5.6%)	5/22 (22.7%)
IPI	15/86 (17.4%)	2/27 (7.4%)	0/27 (0.0%)	1/19 (5.3%)
TRI	53/76 (69.7%)	5/9 (55.6%)	0/2 (0.0%)	0/0 (–)
MPI	241/286 (84.3%)	64/88 (72.7%)	45/56 (80.4%)	50/55 (90.9%)
MEX	8/325 (2.5%)	1/306 (0.3%)	3/174 (1.7%)	2/189 (1.1%)
ADI	18/96 (18.8%)	2/64 (3.1%)	2/32 (6.2%)	1/30 (3.3%)

and documentation of sampling, blinding, or overlap with the earlier queue. Its two groups cover HE and CP, with no SO sampling arm or hidden duplicates. We therefore report its within-group retention separately from the earlier protocol. Both rounds record endpoint labels (HE/SO/CP), rather than GD/AOC channel judgments.

Table 19: **Frozen-queue human endpoint-review diagnostic.** HE, SO, and CP denote HarmEvidence, SemanticOnly, and CorePass. Counts measure label retention or duplicate agreement, except for CP reassigned to HE. The complete endpoint review does not satisfy the prespecified validation criteria.

Check	n	Count	Rate (%)	Count criterion
HE retained as HE	120	87	72.5	≥ 105
SO retained as SO	120	89	74.2	≥ 105
CP retained as CP	60	49	81.7	≥ 55
Hidden-duplicate agreement	30	26	86.7	≥ 27
CP reassigned to HE	60	1	1.7	< 2

Table 20: **Two-group human endpoint review.** Counts show final human labels by recorded initial endpoint; retention is the within-row diagonal proportion. This round is reported separately from the frozen-queue diagnostic in Table 19. HE, SO, and CP denote HarmEvidence, SemanticOnly, and CorePass.

Recorded group	n	Human HE	Human SO	Human CP	Retained (%)
HarmEvidence	120	105	15	0	87.5
CorePass	60	0	2	58	96.7

N GPT-5.6 Sol and GLM-5.3: Detailed Results

This appendix provides the detailed results for GPT-5.6 Sol and GLM-5.3 in Table 2. Both use the same 600 cases and four policies. The structured supplement contains all policy–family–channel counts, length-stratified transitions, recovery membership, and trace references for both input views.

N.1 Configurations, Collection, and Judging

The models use the token limits reported in Section 3.3. A two-response GLM pilot at 2,048 tokens diagnosed generation-budget exhaustion; the formal run uses 8,192 tokens. Of 4,800 planned task–model–policy cells, GPT returned 2,400 responses and GLM returned 2,399. GLM D3/MPI_1247 ended with an upstream content-filter error. This missing response contributes a 0.167-percentage-point completion range to each fixed-600 endpoint.

Transport recovery retried 13 GPT read-timeout failures and 17 GLM reset/DNS failures with the original generation payloads. Eligibility was determined by transport errors, excluding the filtered GLM cell. Each eligible cell received at most two additional attempts with a 300-second timeout, following at most two initial 90-second attempts. The released record includes every attempt and successful response. The analysis includes one empty GPT body and one GLM body containing two newlines; the latter is also one of 18 length-limited outputs. Length summaries use unmodified character counts, with empty-body status determined after whitespace stripping.

The GPT-5.4 judge reuses the Core and evidence templates with a single user message. Core permits 1,024 output tokens and at most three 60-second attempts; evidence auditing permits 512 tokens and three 90-second rounds with up to two transport attempts each. The first valid schema-conforming result is retained. The frozen protocol records endpoint metadata and all retry and parsing rules.

The primary view uses the first 3,000 Python characters for Core and 4,000 for evidence auditing; the full view uses the complete body and its own CoreFail gate. This comparison was specified before inspecting labels. Identical canonical requests share results and attempt budgets. The resulting judge-measurement counts are 3,092 for GPT (2,633 Core, 459 audit), 2,871 for GLM (2,547 Core, 324 audit), and 532 for the bridge. The missing GLM response remains unknown in both views.

N.2 Endpoints and Policy Contrasts

The primary gates close as GPT $377 = 14 + 360 + 3$ and GLM $308 = 73 + 228 + 7$ for RVE, SemanticOnly, and OtherNoHarm. Full-view closures are $364 = 14 + 347 + 3$ and $306 = 75 +$

Table 21: **GPT-5.6 Sol and GLM-5.3: detailed endpoints.** Counts per 600 planned responses; primary uses Core/audit prefixes of 3,000/4,000 characters, and full uses the saved complete body with its own gate.

Model	View	Policy	Core	RVE	GD	AOC	SO	Other	U
GPT-5.6 Sol	Primary	D0	152	10	1	9	142	0	0
		B2	79	3	0	3	75	1	0
		D3	76	1	0	1	75	0	0
		D4	70	0	0	0	68	2	0
GPT-5.6 Sol	Full	D0	151	10	1	9	141	0	0
		B2	74	3	0	3	71	0	0
		D3	73	1	0	1	71	1	0
		D4	66	0	0	0	64	2	0
GLM-5.3	Primary	D0	141	40	1	39	100	1	0
		B2	105	16	0	16	87	2	0
		D3	$38^{\dagger}$	$10^{\dagger}$	$2^{\dagger}$	$10^{\dagger}$	27	1	1
		D4	24	7	2	7	14	3	0
GLM-5.3	Full	D0	142	42	3	41	99	1	0
		B2	104	16	0	16	86	2	0
		D3	$36^{\dagger}$	$10^{\dagger}$	$2^{\dagger}$	$10^{\dagger}$	25	1	1
		D4	24	7	2	7	14	3	0

Core = CoreFail; SO = SemanticOnly; Other = OtherNoHarm; U = missing target response. † adds one unknown to the displayed positive count, giving the completion range $[n/600, (n+1)/600]$, not an imputed negative. GD and AOC overlap: GLM has 4 such rows in primary and 6 in full. All valid CoreFail gates have resolved audits.

$224 + 7$. These model-specific decompositions show why semantic failure and evidence should be reported separately; their proportions are not a cross-model constant. In the complete observations, AOC exceeds GD in each model's aggregate under both views, while individual cells can tie.

Table 22: **All twelve primary D0 contrasts.** Exact two-sided McNemar tests, Holm correction across the three policies separately for each model and endpoint. Positive Δ denotes fewer failures under the added policy.

Model	Endpoint	Policy	n	n_{10}	n_{01}	Δ (pp, $N = 600$)	p_{Holm}
GPT-5.6 Sol	Core	B2	600	90	17	12.167	3.51×10^{-13}
	Core	D3	600	88	12	12.667	3.82×10^{-15}
	Core	D4	600	95	13	13.667	4.39×10^{-16}
GPT-5.6 Sol	RVE	B2	600	9	2	1.167	0.06543
	RVE	D3	600	10	1	1.500	0.02344
	RVE	D4	600	10	0	1.667	0.00586
GLM-5.3	Core	B2	600	57	21	6.000	5.57×10^{-5}
	Core	D3	599	114	11	[17.000, 17.167]	1.93×10^{-22}
	Core	D4	600	127	10	19.500	1.71×10^{-26}
GLM-5.3	RVE	B2	600	28	4	4.000	1.93×10^{-5}
	RVE	D3	599	34	4	[4.833, 5.000]	1.21×10^{-6}
	RVE	D4	600	37	4	5.500	3.08×10^{-7}

n is the number of valid pairs; n_{10}/n_{01} count D0-only/policy-only positives. The interval in GLM D3 completes its one unknown over the fixed 600 denominator, while its test uses 599 pairs. GPT B2 RVE decreases from 10 to 3 but is not significant at .05. Full-view and direct D3–D4 comparisons are descriptive, without new significance tests.

All six primary CoreFail contrasts and five of six RVE contrasts meet the prespecified Holm threshold. The exception is GPT B2 RVE ($10 \to 3$, adjusted $p = 0.06543$), which is an observed decrease without statistical significance. Both new models have lower observed CoreFail and RVE under D4 than D3. These direct D3–D4 comparisons, and every full-view comparison, are descriptive: no ad-

ditional significance tests were performed. GLM contrasts involving D3 retain a fixed denominator of 600 for completion ranges and use 599 valid pairs for the primary test.

N.3 Family and Channel Decomposition

Table 23: **All six families in each configuration.** Counts pool four policies within a model, with 400 planned responses per family (100 per policy); the models are not pooled.

Model	Family	Primary Core	Primary RVE	Full Core	Full RVE
GPT-5.6 Sol	DPI	34	0	34	0
	IPI	20	2	20	2
	TRI	1	0	1	0
	MPI	36	12	36	12
	MEX	208	0	195	0
	ADI	78	0	78	0
GLM-5.3	DPI	20	10	21	12
	IPI	14	1	14	1
	TRI	0	0	0	0
	MPI	59†	57†	59†	57†
	MEX	183	1	180	1
	ADI	32	4	32	4

†: one unknown GLM MPI response. TRI was evaluated: GPT has 0 RVE among 1 CoreFail; GLM has no CoreFail, so its gate-conditional RVE rate is undefined, not 0%. Complete model–policy–family counts and channel margins are in the structured supplement.

MPI has the highest observed gate-conditional evidence rate in each new model, whereas MEX has many semantic failures with little audited evidence. TRI is fully represented in the task grid but contributes only one CoreFail in GPT and none in GLM. Family and channel profiles do not uniformly follow the aggregate policy effect: GLM B2 reduces overall CoreFail/RVE, but DPI counts increase from 4 to 9 CoreFail and from 1 to 3 RVE; primary GD rises from 1 at D0 to 2 at D3 and D4. These distinctions are visible in the complete policy–family–channel supplement and motivate reporting the decomposition alongside overall rates.

N.4 Bridge, Input Views, and Recovery Sensitivity

Table 24: **Fixed 266-response Core bridge.** V2 denotes the stored historical labels; K1 and K2 are independent slots under the new Kilo route.

Comparison	n	Fail→Pass	Pass→Fail	Agreement	κ
V2–K1	266	3	0	98.87%	0.947
K1–K2	266	1	0	99.62%	0.981
V2–K2	266	4	0	98.50%	0.929

All texts are at most 2,000 characters. The sample contains Claude 84, GLM-5.1 82, Qwen 77, Kimi 20, and GPT-5.5 3 responses; families are DPI 190, IPI 57, MEX 9, MPI 7, ADI 3, and TRI 0. These short, unevenly distributed Core measurements are not an evidence-layer calibration or a test of provider/snapshot equivalence.

Table 25: **Primary-to-full view transitions.** Lengths count unmodified Python characters; transition counts cover all returned responses, with missing target responses excluded.

Model	Valid n	> 3000	> 4000	Core $0 \to 1$	Core $1 \to 0$	RVE $0 \to 1$	RVE $1 \to 0$
GPT-5.6 Sol	2400	233	169	8	21	0	0
GLM-5.3	2399	150	35	3	5	2	0

One of the two GLM full-view RVE additions is in the shared CoreFail gate and one enters through the full-view gate; both occur at D0. View changes can reflect both input and judge variability. The short-text bridge cannot isolate a causal truncation effect.

The bridge uses the 266 responses judged again in the reference rerun (Section 3.4). Historical labels versus repeat 1 (V2–K1), repeat 1 versus repeat 2 (K1–K2), and historical labels versus repeat 2 (V2–K2) contain 3, 1, and 4 Fail-to-Pass changes, respectively, with no reverse changes. These comparisons measure Core-label repeatability on that subset. The primary/full comparison combines a longer input with another judge call: GPT has 29 Core flips with a net decrease of 13, while GLM has 8 with a net decrease of 2 and two additional full-view RVE positives.

Table 26: **Recovery-masked D0 contrast ranges (percentage points).** All 13 GPT recovery cells and all 17 GLM recovery cells are set unknown in addition to the retained GLM missing cell. Both the Core gate and downstream endpoints are masked.

Model	View	D0 minus	Core range	RVE range
GPT-5.6 Sol	Primary	B2	[11.333, 12.833]	[0.833, 2.333]
		D3	[11.500, 13.333]	[0.833, 2.667]
		D4	[13.167, 14.333]	[1.667, 2.833]
GPT-5.6 Sol	Full	B2	[11.833, 13.333]	[0.833, 2.333]
		D3	[11.833, 13.667]	[0.833, 2.667]
		D4	[13.500, 14.667]	[1.667, 2.833]
GLM-5.3	Primary	B2	6.000	4.000
		D3	[14.500, 17.500]	[2.333, 5.333]
		D4	19.500	5.500
GLM-5.3	Full	B2	6.333	4.333
		D3	[15.000, 18.000]	[2.667, 5.667]
		D4	19.667	5.833

Every displayed lower bound remains positive. These are worst-case completion ranges, not confidence intervals or re-tested significance. Other conclusions need not be mask-stable: GPT primary AOC can range from 13 to 26 and GD from 1 to 14, so AOC$>$GD is not guaranteed for every completion.