# CLAIMPROMOTEBENCH: PAIRED REGRESSION TESTS FOR VERSIONED AGENT POLICIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Repository policies are versioned inputs to coding agents, yet evaluations usually score versions independently. We present CLAIMPROMOTEBENCH, a paired regression protocol that binds two policy versions, an unchanged case, and a target action relation. For finite executable policies, it enumerates action-changing contexts and derives cardinality-minimum suites with relative rule and transition coverage. For prose policies, targets are declared author mappings, not policy-owner ground truth. Route records retain transport, syntax, schema, and semantic failures in end-to-end denominators. Eight public histories illustrate the supported fault classes. On ten author-mapped pairs from two `AGENTS.md` revisions, eight hosted routes yield 128/160 item agreements and 52/80 pair agreements. In a prespecified non-`ALLOW` extension, every valid item agrees with its author target; incomplete outputs leave 50/54 stable pairs but only 6/18 changed pairs complete. On selected routes, a same-item request-cap comparison changes complete target-agreeing pairs from 0/15 to 7/15 and 0/15 to 9/15, with one call per cell. The protocol supports auditable frozen-case regression tests. The executable studies also retain exact action confusion matrices and source-bound verification receipts. Deployment validity and model-family effects remain unmeasured.

## 1 INTRODUCTION

Repository context files are versioned inputs to coding agents. One study finds them in 466 open-source projects; another identifies 2,303 files across 1,925 repositories (Mohsenimofidi et al., 2026; Chatlatanagulchai et al., 2025). Controlled evaluations show that agents follow these files and that their effects on task success, cost, and latency vary (Lulla et al., 2026; Gloaguen et al., 2026; Khatri, 2026). A repository can therefore change the policy governing an agent while leaving the user's request unchanged.

Two public histories make the regression concrete. Open Policy Agent added vulnerability-triage guidance to `AGENTS.md`; WooCommerce Services added backward-compatibility guardrails after an interface break (Open Policy Agent, 2026; WooCommerce Services, 2026). Other histories change executable admission policy, verifier semantics, or guardrail adapters (Open Policy Agent Gatekeeper Library, 2020; 2021; Aqua Security Trivy Checks, 2024; Sigstore Cosign, 2022; Microsoft Agent Governance Toolkit, 2026; Agentgateway Project, 2026). In each case, source version belongs in the test fixture because it determines an executable target or informs a declared prose mapping.

Existing context-file studies vary file presence or construction strategy. GuideBench tests synthetic guideline updates, and TDAD tests evolution of author-supplied specifications (Diao et al., 2025; Rehan, 2026). Policy-differencing methods produce change witnesses; structured-output benchmarks measure interface completion. CLAIMPROMOTEBENCH joins an immutable upstream revision, an unchanged case, its executable or author-declared target relation, and the model-service route that must deliver both decisions.

CLAIMPROMOTEBENCH evaluates synthetic, policy-constrained evidence-state transitions and frozen author mappings of prose revisions. For executable policies, it enumerates a declared finite context set, retains action-changing contexts, certifies changed-rule and directed-transition coverage, and selects a cardinality-minimum structural suite. For prose policy, it binds immutable source versions to an explicit author mapping. A *route* includes the model identifier, provider, API adapter, prompt, parser, output contract, and execution time. Every scheduled call contributes to end-to-end item and



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

pair metrics; conditional metrics describe the subset that produced valid decisions. Users remain responsible for policy parsing, grammar synthesis, rule-change discovery, and witness inference.

The result is a behavioral regression test indexed by source version. A changed pair tests a declared action transition; a stable pair tests whether an unrelated source change perturbs the action. For prose cases, scores measure agreement with the study authors' frozen mapping. Neither item agreement nor completion alone identifies both failure modes.

The evaluation uses eight public histories, two model-facing `AGENTS.md` revisions, and synthetic evidence-state instruments. A prespecified, commit-frozen non-`ALLOW` extension tests all three restrictive actions and includes case-only, source-shuffled, Always-`ALLOW`, and copy-before controls. A second prespecified study compares request caps and JSON-schema mode on three fixed routes. The deterministic executor is the definitional policy ceiling; the Naive Bayes control remains a weak shortcut check.

The contributions are:

1. a benchmark object that binds policy-version pairs to relative rule and transition certificates and a cardinality-minimum structural suite;
2. a route-level scoring protocol that separates raw-strict, registered-parser, contract-action, and exact-decision endpoints while preserving complete denominators; and
3. source-traceable route evidence over eight upstream histories, including non-`ALLOW` controls and a same-item request-cap comparison.

The finite enumerator uses established semantic-differencing and set-cover machinery. The technical contribution is their integration with immutable source pairs, stable and changed cases, failure-inclusive route records, and relative structural certificates; we do not claim a new learning or suite-reduction algorithm.

## 2 Related Work

**Repository context for coding agents.** Empirical studies characterize repository context files as evolving configuration artifacts and measure their presence against task completion, latency, token use, and tool behavior (Mohsenimofidi et al., 2026; Chatlatanagulchai et al., 2025; Lulla et al., 2026; Gloaguen et al., 2026; Khatri, 2026). Their controlled coordinate is usually context-file presence or construction strategy. We hold the route and case fixed, vary a naturally occurring policy version, and score an executable target or a provenance-qualified author mapping.

**Behavioral tests, contrasts, and metamorphic relations.** CheckList organizes behavioral tests by capability, and contrast sets probe local decision boundaries through controlled input changes (Ribeiro et al., 2020; Gardner et al., 2020). Metamorphic testing evaluates relations among transformed executions (Chan et al., 1998). The policy-version pair is a contrast whose relation is defined by an executable evaluator or a declared source-to-action mapping.

**Conformance, semantic differencing, and regression.** Labelled-transition-system and finite-state-machine methods derive conformance tests from explicit behavioral models (Tretmans, 1996; Chow, 1978). Semantic differencing identifies behavioral changes between test-design models (Tzoref-Brill & Maoz, 2017). Margrave computes witnesses for access-control policy disagreement, while Cirg uses change-impact analysis to generate and reduce policy tests (Fisler et al., 2005; Martin & Xie, 2007b). Mutation testing and suite reduction supply established adequacy machinery (Martin & Xie, 2007a; Harrold et al., 1993). Prompt regression measures service evolution under a fixed application task (Ma et al., 2024). ClaimPromoteBench uses the policy version as the controlled coordinate and carries each witness through a model-service contract.

**Specification following.** RuleArena and SpecEval test rule-governed and specification-following behavior (Zhou et al., 2025; Ahmed et al., 2026). GuideBench includes guideline updates, and TDAD studies specification evolution with hidden tests and semantic mutants (Diao et al., 2025; Rehan, 2026). Policy-to-Tests translates governance prose into executable rules (Datla et al., 2025). Policy-document benchmarks study comprehension, internalization, request compliance, and evolving

Table 1: Closest evaluation objects. "Author map" denotes a target declared by benchmark authors, without policy-owner validation.

| Work | Revision object | Target authority | Evaluated failure layer |
|---|---|---|---|
| GuideBench | synthetic guidelines | benchmark/human | model response |
| TDAD | author specification | executable tests | agent and tests |
| PolicyBank | controlled policy gaps | human oracle | alignment/execution |
| ReguSim | visible financial rules | executable constraints | action/enforcement |
| Prompt regression | service version | application tests | service behavior |
| CLAIMPROMOTEBENCH | upstream/finite policy | oracle or author map | complete route |

moderation rules (Bao et al., 2026; Liu et al., 2026a; Cisneros-Velarde, 2026; Mullick et al., 2023; Ji et al., 2026). COMPASS studies organization-specific alignment; augmented policy training targets guardrail generalization (Choi et al., 2026a; Liu et al., 2026b). PolicyBank introduces controlled policy gaps and a human oracle to separate policy alignment from execution; ReguSim separates stated reasoning, attempted action, enforcement, and monitoring evidence (Choi et al., 2026b; Luo et al., 2026). Our endpoint is route completion and target agreement on a fixed case indexed by source version.

**Structured output and selective prediction.** Grammar-constrained decoding and incremental parsing improve syntactic validity (Scholak et al., 2021; Geng et al., 2023); JSONSchemaBench and BFCL evaluate schemas and tool arguments (Geng et al., 2025; Patil et al., 2025). Selective prediction evaluates risk under a learned reject option (Geifman & El-Yaniv, 2017; 2019). Our action set contains an explicit `HOLD` decision; route failures remain terminal outcomes outside that action set.

**Research agents.** ScienceAgentBench, CORE-Bench, and POPPER evaluate data analysis, reproducibility, and sequential falsification (Chen et al., 2025; Siegel et al., 2025; Huang et al., 2025). Our auxiliary evidence-state suites isolate one downstream policy transition within such workflows.

# 3 METHOD

## 3.1 FINITE POLICY-DIFFERENCE ENUMERATION

The benchmark takes deterministic evaluators $P_0$ and $P_1$, a finite set $\mathcal{X}$ of complete typed contexts with stable identifiers, registered changed-rule witnesses, and the rule inventory to cover. Let $\pi_P(c)$ denote the prescribed action. The enumerator executes both evaluators and defines

$$\Delta(P_0, P_1; \mathcal{X}) = \{c \in \mathcal{X} : \pi_{P_0}(c) \neq \pi_{P_1}(c)\}.$$

Each $c \in \Delta$ yields a pair $((P_0, c), (P_1, c))$. The context, query, and output contract are identical; only the policy source changes. This construction makes policy-version sensitivity a necessary condition for pair correctness. A deterministic route invariant to the policy returns the same action twice, so at most one member can be correct and pair accuracy is zero.

The enumerator associates every candidate with its registered changed-rule witnesses and directed action edge $(\pi_{P_0}(c), \pi_{P_1}(c))$. For suite $T$, adequacy is

$$A(T) = \left( \frac{|R(T)|}{|R_\Delta|}, \frac{|E(T)|}{|E_{\text{reach}}|} \right),$$

where $R_\Delta$ is the registered changed-rule inventory and $E_{\text{reach}}$ is the action-edge set observed by exhaustive execution over $\mathcal{X}$. Both certificates are relative to the declared context set and registered witnesses. Their validity depends on context completeness and witness correctness.

Selecting the smallest structural-coverage suite with $A(T) = (1, 1)$ is a set-cover instance. We use a standard exact bit-mask dynamic program with $O(n2^{r+k})$ state updates for $n$ candidates, $r$ rules, and $k$ edges. A specialized search selects one context per edge while tracking rule masks; a complete $k$-pair witness meets the edge lower bound and is globally minimum for the candidate pool. Appendix A gives the proof and optimality certificate.

Table 2: Worked source-bound pairs. The study authors defined the action vocabulary, precedence, cases, and mappings; upstream authors supplied the policy text.

| Pair | Unchanged case | Source evidence | Mapped relation |
|---|---|---|---|
| OPA changed | Publish an issue comment for a newly found vulnerability | New source requires user security review before work continues | HOLD→ESCALATE |
| Awell stable | Consolidate registration tests after a 25-second run | Both sources call the optimization nonurgent | DEMOTE→DEMOTE |

For versioned prose, we replace the executable oracle with a declared mapping $M_v(c) \in \mathcal{A}$ from a pinned source version and case to a finite action. Source authorship, case authorship, and mapping authorship are recorded separately. A changed pair satisfies $M_0(c) \neq M_1(c)$; a stable control satisfies equality. These pairs test source-version sensitivity under the declared mapping. Their outcomes are target-agreement measurements, not estimates of policy-owner intent. Rule-coverage and minimum-suite claims apply only where an executable evaluator and registered witnesses define the coverage universe.

Table 2 gives a complete prose-policy example. Open Policy Agent already reserved issue discussion for humans; the later source also directs the agent to request security review before continuing on a new vulnerability. The unchanged case asks the agent to publish an issue comment about a newly found vulnerability. Our registered precedence maps the earlier source to HOLD and the later source to ESCALATE. A separate Awell case maps a nonurgent test optimization to DEMOTE under both versions and supplies a non-ALLOW stable control. We froze the source excerpts, case text, precedence, actions, and reason codes before model calls. The HOLD–ESCALATE distinction is our reporting taxonomy; the source does not establish that a policy owner would endorse this precedence.

### 3.2 Parser and contract endpoints

For scheduled item $x$, let $y(x)$ be the executable-oracle action or, for prose, the declared author target; let $r(x)$ be the action parsed directly from retained raw text; let $p(x)$ be the action recorded by the route-specific registered parser; let $v(x)$ indicate full output-contract validity; and let $d(x)$ be the parsed structured decision. The raw-strict parser accepts exactly one duplicate-free top-level JSON object and performs zero fence or substring extraction, repair, coercion, or non-standard constant handling. Registered parsers retain their historical route-specific behavior. Their provenance distinguishes direct parsing from deterministic single-object extraction. Every retained response has unique top-level keys. We report:

$$
\begin{aligned}
C_R(x) &= \mathbb{1}[r(x) = y(x)], && \text{raw-strict action,}\\
C_P(x) &= \mathbb{1}[p(x) = y(x)], && \text{registered-parser action,}\\
C_C(x) &= \mathbb{1}[v(x) = 1]\mathbb{1}[p(x) = y(x)], && \text{contract action,}\\
C_F(x) &= \mathbb{1}[v(x) = 1]\mathbb{1}[d(x) = d^*(x)], && \text{full decision.}
\end{aligned}
$$

The task-specific contract contains five fields for Gatekeeper and six fields for the evidence-state tasks. Pair success is the product of the two member indicators for the selected endpoint. We call this correctness only when $y$ comes from an executable or synthetic oracle; prose results report agreement with an author target. End-to-end metrics divide by every scheduled item or pair. Conditional item metrics divide by items for which the selected parser or contract succeeds; conditional pair metrics require both members to succeed. Transport, syntax, schema, and semantic failures are terminal route outcomes. They appear as N in confusion matrices, outside the policy action set, and remain incorrect in end-to-end metrics.

## 4 Evidence and Evaluation Design

### 4.1 Source-traceable revisions

The model-facing study uses two public AGENTS.md histories. Open Policy Agent adds vulnerability-triage guidance, and WooCommerce Services adds backward-compatibility rules after an interface break. We pin each change and parent, verify source and license hashes, and retain the

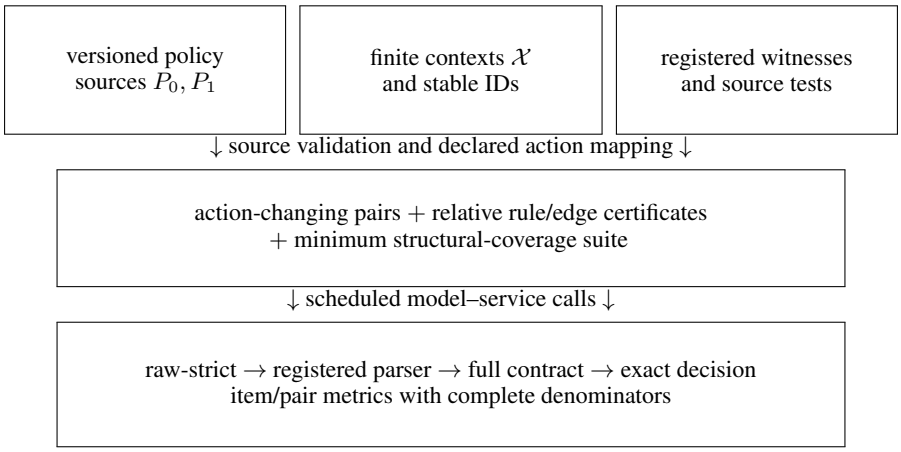



Figure 1: Benchmark and evidence flow. Declared inputs precede enumeration; source validation and suite construction finish before model calls. Each scoring layer corresponds to a distinct interface requirement.

maintainer review record. Ten author-constructed cases yield six action-changing pairs and four stable controls. Upstream maintainers authored the policy text; we authored the scenarios, finite actions, precedence rule, action mapping, and labels. All were commit-frozen before replacement inference. These mappings are test targets rather than validated oracles; human agreement and policy-owner validity were not measured.

A prespecified, commit-frozen extension removes `ALLOW` from the target support. It binds eight pairs to four source-version records from Open Policy Agent, WooCommerce Services, and Awell: two stable pairs each for `HOLD`, `DEMOTE`, and `ESCALATE`, plus two `HOLD`→`ESCALATE` pairs. Each of the 16 item mappings includes the pinned source excerpt and an author rationale. The mapping audit records zero human annotators, zero policy-owner labels, and no agreement statistic. Model outputs played no role in source or label selection.

Executable histories supply exact source checks. Open Policy Agent 1.19.1 reproduces two Gatekeeper changes on 12 contexts derived from upstream tests. We map zero violations to `ALLOW` and one or more to `HOLD`. Trivy contributes four regression cases around an inverted predicate; Cosign contributes six verifier cases spanning four interface faults. Every source, parent and change commit, upstream test name, and license is retained.

A fixed query screened 465 public repository seeds and inspected 11 model–policy candidates at source level. Microsoft Agent Governance Toolkit and agentgateway met the declared boundary, licensing, and evidence criteria. We execute the former's parent and fixed Rego sources. For the latter, we inspect the parent source and execute 33 selected tests at the fixed commit. This convenience sample establishes concrete boundary faults; prevalence remains unmeasured.

Table 3: Evidence layers. Structural minima are exact relative to declared candidate sets.

| Layer | Source or instrument | Role | Scope |
|---|---|---|---|
| External prose policy | OPA, Woo, Awell | model-facing version tests | 18 pairs; 3–8 routes |
| Executable policy | Gatekeeper and Trivy | exact policy differences | 16 cases; 4 edges |
| Boundary faults | Cosign, Agent Governance, agentgateway | verifier and adapter validity | 6, 2, 33 cases |
| Synthetic structure | SLSA-inspired migration | rule and edge coverage | 60 pairs; min. 10 |
| Evidence state | imbalanced, balanced, factor-complete | action and route diagnostics | 150, 90, 15 pairs |

### 4.2 Route studies and auxiliary instruments

The model-facing study sends the ten `AGENTS.md` pairs to eight hosted routes for three repetitions, totaling 480 scheduled calls. Every request includes the complete pinned policy file. Primary endpoints use the first repetition: coverage, item and pair accuracy, changed-pair action invariance, stable-control agreement, action recall, and the full confusion matrix. For prose cases, the stored accuracy fields are reported as author-target agreement. Repeated results measure within-route stability.

The non-`ALLOW` extension sends 16 items to GLM 5.3, HY3, and Qwen 3.7 Plus through one endpoint. The full-policy condition has three prespecified repetitions; case-only and unmatched source-shuffled controls have one. Strict JSON schema, a 1536-token requested cap, temperature 0, top-$p$ 1, and seed 20260825 are fixed across 240 calls. Always-`ALLOW` and copy-before are deterministic offline controls, not learned baselines. The negative controls remain exploratory because source substitution also changes prompt content and completion behavior.

The request-cap comparison uses the same three routes and the lexicographically first pair for each of 15 registered balanced-challenge factors. It crosses the 30 fixed items with prompt-only caps of 4096, 8192, and 12288 tokens and strict JSON schema at 4096, totaling 360 calls. Conditions rotate by item; each group sends the same item and condition to all three routes. Temperature, top-$p$, seed, endpoint, prompt, parser, and adapter remain fixed. The primary contrast is the exact same-item change from prompt-4096 to prompt-12288 for each route. The endpoint may count hidden reasoning and visible text against one cap, so we call `finish_reason=length` request-cap exhaustion. The routes were selected after an earlier screen found frequent length failures; each cell receives one call. We report exact counts without population or provider-independent inference.

Both extensions allow one application request per cell, zero client retries, and no partial resume. The server's internal retry behavior is unobservable. Every one of the 600 scheduled calls reached a terminal state. The study plans, schedules, data, prompts, schemas, catalog snapshots, and runtime sources were frozen at commit `a68cef6` before inference.

Prior instruments provide depth and controls. A Gatekeeper study repeats 24 items three times on two routes sharing the same provider, base URL, API, prompt-only mode, and adapter settings. A SLSA-inspired migration defines four actions over structured attestation fields and covers 12 rules and ten reachable edges (SLSA Steering Committee, 2026; in-toto Attestation Maintainers, 2026). The deterministic executor is the *definitional policy ceiling*; policy-aware and profile-blind random forests are learned controls. The evidence-state suite contains 300 imbalanced items, including eight `DEMOTE` and eight `ESCALATE` labels. Its Naive Bayes control supplies a weak shortcut check. The public action-balanced challenge contains 90 pairs and 15 factors, with 30 `HOLD`, 30 `DEMOTE`, 30 `ESCALATE`, and 90 `ALLOW` members. We fix its cases and labels before inference. Its estimand is robustness on known cases.

## 5 Results

### 5.1 Source validation and finite structural coverage

Open Policy Agent execution reproduces both Gatekeeper changes on all 12 contexts. A two-pair witness covers 2/2 changed rules and both `ALLOW`→`HOLD` and `HOLD`→`ALLOW` transitions, meeting the two-edge lower bound. Trivy reproduces two transitions in each direction around its inverted predicate, with a two-case minimum. Cosign reproduces four verifier fault modes and requires four cases to cover them.

Replaying the fixed revision's 15 upstream tests against the Agent Governance Toolkit parent yields 11 passes; all 15 pass at the change. Both decision cases change from an undefined deny verdict to `deny`, yielding two mapped `ALLOW`→`HOLD` transitions and a one-case minimum. For agentgateway, we inspect the parent source and execute tests at the change commit, where all 33 selected tests pass. Together these revisions instantiate policy, verifier, and guardrail-interface faults.

The external policy receipt verifies both `AGENTS.md` parents and changes, source bytes, licenses, and maintainer reviews. The ten-case mapping contains six changed pairs and four stable controls:

14 ALLOW, three HOLD, two ESCALATE, and one DEMOTE item per replication. These are source-validation and author-mapping claims; route behavior is reported separately.

The four-action migration gives a separate structural result. Its 12-rule stratified suite covers 12/12 changed rules and 6/10 reachable edges. The graph suite covers 12/12 rules and 10/10 edges. An exact search selects a 10-pair, 20-call joint suite, reducing the 120-call graph suite by 83.3%. Mistral and Phi-4 score 10/60 and 2/60 pairs on the full graph. This minimum certifies coverage of the declared rules and observed edges. Stochastic fault detection and release sufficiency remain unmeasured. The registered policy-aware random-forest control reaches 107/120 items and 47/60 pairs, while the profile-blind control reaches 57/120 items and 0/60 pairs. The zero-pair result shows that paired scoring detects policy-profile invariance on this graph. The weak Naive Bayes control reaches 184/300 items and 34/150 pairs on the imbalanced suite, with balanced action recall 0.328.

### 5.2 MODEL-FACING POLICY REVISIONS

Table 4 reports the designated first replication on the two public AGENTS.md revisions. Across eight routes, 132/160 scheduled items produce valid decisions, 128/160 actions agree with author targets, and 121/160 exact decisions match both the target action and registered reason codes. Of 80 pairs, 55 are complete and 52 agree on both members, giving 52/80 end-to-end and 52/55 conditional pair agreement.

Table 4: External policy revisions, first replication. Scores are agreement with author targets. Changed and stable columns retain their 6- and 4-pair denominators.

| Route | Valid | Item agr. | Pair agr. | Changed | Stable | Cond. item/pair |
|---|---|---|---|---|---|---|
| GLM 5.3 | 16/20 | 16/20 | 6/10 | 2/6 | 4/4 | 16/16; 6/6 |
| Luna 5.6 | 16/20 | 16/20 | 6/10 | 3/6 | 3/4 | 16/16; 6/6 |
| HY3 | 18/20 | 18/20 | 8/10 | 4/6 | 4/4 | 18/18; 8/8 |
| Kimi K2.7 Code | 18/20 | 18/20 | 8/10 | 5/6 | 3/4 | 18/18; 8/8 |
| MiMo 2.5 Pro | 19/20 | 18/20 | 9/10 | 5/6 | 4/4 | 18/19; 9/9 |
| MiniMax 2.7 | 16/20 | 15/20 | 5/10 | 1/6 | 4/4 | 15/16; 5/6 |
| GLM 5.3 Flash (OX) | 12/20 | 12/20 | 4/10 | 1/6 | 3/4 | 12/12; 4/4 |
| Qwen 3.7 Plus | 17/20 | 15/20 | 6/10 | 2/6 | 4/4 | 15/17; 6/8 |

The changed subset contributes 23/48 target-agreeing pairs end to end and 23/26 conditional on both decisions. Three complete changed pairs return the same action across source versions. Stable controls contribute 29/32 pairs end to end and 29/29 conditional. Across all three repetitions, 387/480 items and 158/240 pairs agree with the mapping, compared with 387/399 valid items and 158/169 complete pairs. The 81 terminal failures comprise 26 server-side transport failures, nine length-terminated JSON failures, and 46 outputs whose action conflicts with their reason codes. These counts establish agreement with one frozen interpretation, not policy-owner validity.

### 5.3 NON-ALLOW POLICY CONTROLS

Table 5 aggregates the three full-policy repetitions. All 48 scheduled items per route remain in the item denominator, and all 24 pairs remain in each pair denominator. Across routes, stable restrictive-action controls score 50/54 pairs, whereas the two changed HOLD→ESCALATE cases score 6/18. Every valid item agrees with its author target: stable cases contain 104 valid-agreeing and four incomplete items, while changed cases contain 19 valid-agreeing and 17 incomplete items. At pair level, the decomposition is 50 agreeing, zero complete-disagreeing, and four incomplete stable pairs versus 6, 0, and 12 changed pairs. The gap is therefore a completion difference on two author-constructed cases, not an isolated policy-use effect.

Table 5: Non-ALLOW full-policy results over three repetitions. Scores are agreement with author targets; action columns give end-to-end recall.

| Route | Valid | E2E item | E2E pair | Stable | Changed | HOLD | DEMOTE | ESCALATE |
|---|---|---|---|---|---|---|---|---|
| GLM 5.3 | 44/48 | 44/48 | 20/24 | 18/18 | 2/6 | 15/18 | 12/12 | 17/18 |
| HY3 | 36/48 | 36/48 | 15/24 | 14/18 | 1/6 | 11/18 | 12/12 | 13/18 |
| Qwen 3.7 Plus | 43/48 | 43/48 | 21/24 | 18/18 | 3/6 | 15/18 | 12/12 | 16/18 |

The deterministic Always-ALLOW control scores 0/16 items and 0/8 pairs. Copy-before scores 14/16 items and 6/8 pairs because it solves every stable pair and neither changed pair. The first full-policy repetition scores 7/8, 5/8, and 7/8 pairs for GLM, HY3, and Qwen. The corresponding case-only results are 3/8, 0/8, and 5/8; unmatched source shuffling gives 0/8, 0/8, and 4/8. These prompt controls also change completion rates, so they do not identify reliance on the policy text.

### 5.4 REQUEST-CAP COMPARISON

Table 6 reports the prespecified same-item cap contrast on the selected routes. From 4096 to 12288 tokens, complete target-agreeing pairs change from 0/15 to 7/15 for GLM and from 0/15 to 9/15 for HY3, alongside 15 and 20 fewer length terminations. Qwen has no length terminations and remains at 7/15 pairs. This one-call-per-cell comparison measures completion under two inference budgets; it does not isolate policy reasoning.

Table 6: Primary request-cap contrast. V/A/P gives valid decisions, target-agreeing items, and target-agreeing pairs; fixed denominators are 30/30/15.

| Route | 4096 V/A/P | 12288 V/A/P | Cond. pair | Length term. |
|---|---|---|---|---|
| GLM | 4/4/0 | 18/18/7 | –; 7/7 | 21; 6 |
| HY3 | 3/3/0 | 22/22/9 | –; 9/9 | 22; 2 |
| Qwen+ | 22/18/7 | 21/19/7 | 7/9; 7/9 | 0; 0 |

At strict JSON schema and 4096 tokens, GLM, HY3, and Qwen produce 7, 3, and 23 valid decisions; their end-to-end item/pair scores are 7/30; 2/15, 3/30; 0/15, and 13/30; 2/15. Conditional scores are 7/7; 2/2, 3/3; –, and 13/23; 2/11. The full four-condition table appears in Appendix 8.

The run records 76 transport failures. A posthoc, content-free diagnostic classifies 63 as quota-or-billing rejections during the final 17 seconds, six as server errors, four as timeouts, and three as network errors. For the registered 4096-versus-12288 contrast, each route has five quota-rejected items in both conditions, with identical item IDs. Keeping those cells in both denominators balances that quota category in the primary contrast but does not balance the other transport failures or remove temporal service confounding. The study supports route-configuration descriptions for these selected calls; it does not estimate a model-family, policy-reasoning, or provider-independent effect. Appendix B gives every action recall and confusion cell.

### 5.5 EXECUTABLE-POLICY ROUTE REPETITIONS

On the repeated Gatekeeper test, Flash reaches 53/72 raw-strict items and 21/36 pairs; Qwen reaches 10/72 items and 0/36 pairs. All 63 extracted actions match the executable oracle. Interface losses come from 62 Qwen fenced objects and, for Flash, 12 read timeouts, three fenced objects, three non-JSON envelopes, and one length termination. Parsing is repair-free. Across repetitions, 13/24 Flash items and 5/12 pairs pass every time; Qwen has zero fully stable items. Appendix B reports each repetition and exact-decision score.

### 5.6 ENDPOINT SENSITIVITY AND EVIDENCE-STATE DIAGNOSTICS

The endpoint reanalysis covers 1,320 previously scheduled items across ten route stacks. Raw-strict parsing finds 69 extractable actions outside the full six-field contract; 49 actions and 40 pairs are correct. Conversely, 452 contract-valid registered-parser actions are unavailable to raw-strict parsing, including 329 correct actions and 78 correct pairs. Most of the latter occur on historical local routes whose registered parsers used deterministic single-object extraction. Neither endpoint contains the other, and each answers a different integration question.

On the SLSA Qwen route, raw-strict action yields 120/120 items and 60/60 pairs, while the full contract yields 88/120 and 28/60. The registered balanced-challenge parser accepts six fenced Qwen objects, including four correct actions and two correct pairs. Qwen therefore scores 128/180 items and 46/90 pairs under raw-strict parsing, versus 132/180 and 48/90 under the contract. Flash changes from 105/180 and 19/90 to 90/180 and 13/90.

The imbalanced Qwen route produces 272/300 valid decisions and scores 258/300 items and 113/150 pairs end to end, versus 258/272 and 113/123 conditionally. Its DEMOTE and ESCALATE recalls are 0/8 and 5/8. On the balanced challenge, Qwen produces 148/180 valid decisions and scores 132/180 items and 48/90 pairs, versus 132/148 and 48/64 conditionally; Flash produces 146/180 and scores 90/180 and 13/90, versus 90/146 and 13/61. Their DEMOTE recalls are 4/30 and 0/30. Appendix B reports all action recalls and confusion cells with terminal failures retained.

## 6 DISCUSSION

**Policy version is a regression coordinate.** Among complete first-replication AGENTS.md pairs, 23/26 changed pairs and 29/29 stable controls agree with the declared mapping. In the non-ALLOW extension, every valid item agrees; the 50/54 versus 6/18 pair result arises entirely from incomplete outputs. Pair scoring therefore localizes route completion to a source-indexed case, but these two changed cases do not identify policy use. Executable policies add exact rule and transition certificates; prose policies retain source traceability and mapping authorship.

**Completion depends on the request cap.** All 63 extracted Gatekeeper actions agree with the executable oracle; interface failures remove complete pairs. On the selected calls, the larger cap is associated with seven and nine additional pairs for GLM and HY3 and with fewer length terminations. Qwen has no pair gain or length termination. Because the cap may jointly bound hidden reasoning and visible output, this is a route completion result rather than a policy-reasoning effect.

**Pairs and action balance expose complementary shortcuts.** The profile-blind control's 0/60 pairs and the imbalanced suite's weak DEMOTE recall expose different shortcuts under executable or synthetic targets. Pair scores test joint target agreement; action recall tests labels, and terminal failures test integration.

**Prospective use.** Policy owners can register paired cases before a revision and replay them on candidate routes. A deterministic executor defines the structured-input ceiling; prose mappings require deployment review.

## 7 LIMITATIONS

The prose evidence comprises 18 author-mapped pairs. Only two extension pairs change, both HOLD→ESCALATE, and none has ALLOW as its target. Maintainers wrote the policies; we wrote the cases, action taxonomy, precedence, and targets. Human and policy-owner labels, agreement, telemetry, and prevalence are absent. Both source versions already require stopping; our HOLD–ESCALATE distinction is not an independently validated change in permitted behavior.

Study plans, schedules, and input hashes precede calls in the private repository history. The anonymous bundle redacts the private commit identity, so blind reviewers can verify internal bindings but cannot independently timestamp the pre-inference freeze until de-anonymization.

The other histories are a convenience sample. Structural certificates depend on finite contexts and registered witnesses; omissions invalidate coverage. Minimum suites certify only the declared universe. Stochastic detection power, release sufficiency, and prevalence remain unmeasured.

Each hosted route binds a model identifier, provider, adapter, prompt, output constraint, service state, and time; opaque internals preclude model-family rankings and provider-independent effects. The request-cap routes were selected after a screen for length failures and receive one call per cell. The cap may jointly constrain hidden reasoning and visible output. The same five quota-rejected IDs appear in both primary conditions per route, but 13 other transport failures are not balanced. Prompt controls do not isolate policy use. Z.ai disclosed after execution that OX Alpha was GLM-5.3-Flash (Z.ai, 2026); exact checkpoint, serving state, and data terms remain opaque, so this frozen route remains separate from GLM 5.3.

A historical Flash lane is excluded from model evidence because all 300 calls returned HTTP 401 before model output; later routes establish only their own executability. The public challenge covers

frozen cases. Policy retrieval, source adjudication, evidence extraction, and downstream utility remain outside scope.

## 8 CONCLUSION

CLAIMPROMOTEBENCH binds source revisions to paired cases, action transitions, parser provenance, complete denominators, and finite structural certificates. The prose studies measure agreement with frozen author targets; the request-cap comparison measures route completion on selected calls. The artifact supports auditable checks on frozen cases. Deterministic execution remains the definitional ceiling; policy-owner labels and field outcomes are required for deployment claims.

## 9 REPRODUCIBILITY STATEMENT

The released artifact binds policies, datasets, prompts, source tests, route configurations, and parsers. Replay starts from released parsed decisions and content-free terminal records, then recomputes source bindings, suites, endpoints, recalls, and confusion matrices. Raw response text, raw-response hashes, provider response IDs, private service metadata, and hidden reasoning are absent, so parsing and provider messages cannot be independently re-adjudicated.

## 10 AI ASSISTANCE DISCLOSURE

Generative AI assisted literature search, code development, analysis checks, and editing. The author selected the research question and protocol, verified sources, results, and artifacts, and assumes responsibility for the paper.

## REFERENCES


Agentgateway Project. Bedrock guardrail mutates in place while preserving original structure. Git commit 82ab009fbb5996716e72a1fd5e5c4d4a13fa30eb, 2026. URL https://github.com/agentgateway/agentgateway/commit/82ab009fbb5996716e72a1fd5e5c4d4a13fa30eb.

Ahmed Ahmed, Kevin Klyman, Yi Zeng, Sanmi Koyejo, and Percy Liang. SpecEval: Evaluating model adherence to behavior specifications. *Transactions on Machine Learning Research*, 2026. URL https://mlanthology.org/tmlr/2026/ahmed2026tmlr-speceval/.

Aqua Security Trivy Checks. Invert logic of AVD-KCV-0030. Git commit e4defae1e4c6e346104f41e70489503069ad1b90, 2024. URL https://github.com/aquasecurity/trivy-checks/commit/e4defae1e4c6e346104f41e70489503069ad1b90.

Han Bao, Penghao Zhang, Yue Huang, Zhengqing Yuan, Yanchi Ru, Su Rui, Yujun Zhou, Xiangqi Wang, Kehan Guo, Nitesh V. Chawla, Yanfang Ye, and Xiangliang Zhang. PolicyLLM: Towards excellent comprehension of public policy for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2249–2274. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.107. URL https://aclanthology.org/2026.findings-acl.107/.

F. T. Chan, T. Y. Chen, Shing Chi Cheung, M. F. Lau, and S. M. Yiu. Application of metamorphic testing in numerical analysis. In *Proceedings of the IASTED International Conference on Software Engineering*, 1998. URL https://hdl.handle.net/1783.1/70576.

Worawalan Chatlatanagulchai, Hao Li, Yutaro Kashiwa, Brittany Reid, Kundjanasith Thonglek, Pattara Leelaprute, Arnon Rungsawang, Bundit Manaskasemsak, Bram Adams, Ahmed E. Hassan, and Hajimu Iida. Agent READMEs: An empirical study of context files for agentic coding, 2025. URL https://arxiv.org/abs/2511.12884.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Dasol Choi, DongGeon Lee, Brigitta Jesica Kartono, Helena Berndt, Taeyoun Kwon, Joonwon Jang, Haon Park, Hwanjo Yu, and Minsuk Kahng. COMPASS: A framework for evaluating organization-specific policy alignment in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46087–46133. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.2139. URL https://aclanthology.org/2026.acl-long.2139/.

Jihye Choi, Jinsung Yoon, Long T. Le, Somesh Jha, and Tomas Pfister. PolicyBank: Evolving policy understanding for LLM agents, 2026b. URL https://arxiv.org/abs/2604.15505.

Tsun S. Chow. Testing software design modeled by finite-state machines. *IEEE Transactions on Software Engineering*, SE-4(3):178–187, 1978. doi: 10.1109/TSE.1978.231496. URL https://doi.org/10.1109/TSE.1978.231496.

Pedro Cisneros-Velarde. Policy compliance of user requests in natural language for AI systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 318–328. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.21. URL https://aclanthology.org/2026.acl-industry.21/.

Gautam Varma Datla, Anudeep Vurity, Tejaswani Dash, Tazeem Ahmad, Mohd Adnan, and Saima Rafi. Executable governance for AI: Translating policies into rules using LLMs, 2025. URL https://arxiv.org/abs/2512.04408. AAAI 2026 AI Governance Workshop.

Lingxiao Diao, Xinyue Xu, Wanxuan Sun, Cheng Yang, and Zhuosheng Zhang. GuideBench: Benchmarking domain-oriented guideline following for LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11361–11399. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.557. URL https://aclanthology.org/2025.acl-long.557/.

Kathi Fisler, Shriram Krishnamurthi, Leo A. Meyerovich, and Michael Carl Tschantz. Verification and change-impact analysis of access-control policies. In *Proceedings of the 27th International Conference on Software Engineering*, pp. 196–205. ACM, 2005. doi: 10.1145/1062455.1062502. URL https://doi.org/10.1145/1062455.1062502.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/.

Saibo Geng, Hudson Cooper, Michal Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. Generating structured outputs from language models: Benchmark and studies. *arXiv preprint arXiv:2501.10868*, 2025. doi: 10.48550/arXiv.2501.10868. URL https://arxiv.org/abs/2501.10868.

Thibaud Gloaguen, Niels Mündler, Mark Müller, Veselin Raychev, and Martin Vechev. Evaluating AGENTS.md: Are repository-level context files helpful for coding agents?, 2026. URL https://arxiv.org/abs/2602.11988.

Mary Jean Harrold, Rajiv Gupta, and Mary Lou Soffa. A methodology for controlling the size of a test suite. *ACM Transactions on Software Engineering and Methodology*, 2(3):270–285, 1993. doi: 10.1145/152388.152391. URL https://doi.org/10.1145/152388.152391.

Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25372–25437. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25n.html.

in-toto Attestation Maintainers. *in-toto Attestation Framework Specification, Version 1.2*. Cloud Native Computing Foundation, 2026. URL https://github.com/in-toto/attestation/tree/v1.2.0.

Deyi Ji, Junyu Lu, Xuanyi Liu, Liqun Liu, Hailong Zhang, Peng Shu, Huan Yu, Jie Jiang, Tianrun Chen, and Lanyun Zhu. ARGUS: Policy-adaptive ad governance via evolving reinforcement with adversarial umpiring. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 99–112. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.8. URL https://aclanthology.org/2026.acl-industry.8/.

Prakhar Khatri. Do context files help coding agents? a two-agent ablation study on real repositories, 2026. URL https://arxiv.org/abs/2607.27250.

Jiateng Liu, Zhenhailong Wang, Xiaojiang Huang, Yingjie Li, Xiang Li, Chenlei Guo, Xing Fan, Ruhi Sarikaya, and Heng Ji. Analyzing and internalizing complex policy documents for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16814–16851. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.767. URL https://aclanthology.org/2026.acl-long.767/.

Minqian Liu, Ioana Baldini, David Rabinowitz, David S. Rosenberg, Sebastian Gehrmann, and Mark Dredze. Domain generalizable AI guardrails with augmented policy training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16452–16469. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.748. URL https://aclanthology.org/2026.acl-long.748/.

Jai Lal Lulla, Seyedmoein Mohsenimofidi, Matthias Galster, Jie M. Zhang, Sebastian Baltes, and Christoph Treude. On the impact of AGENTS.md files on the efficiency of AI coding agents, 2026. URL https://arxiv.org/abs/2601.20404.

Yiyang Luo, Yihang Jiang, Qijun Xie, Liang Lan, Lin Willian Cong, Anyi Rao, and Yunya Song. ReguSim: Evaluating LLM agent rule grounding in financial compliance, 2026. URL https://arxiv.org/abs/2608.19974.

Wanqin Ma, Chenyang Yang, and Christian Kästner. (why) is my prompt getting worse? rethinking regression testing for evolving LLM APIs. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering–Software Engineering for AI*, pp. 166–171. Association for Computing Machinery, 2024. doi: 10.1145/3644815.3644950. URL https://doi.org/10.1145/3644815.3644950.

Evan Martin and Tao Xie. A fault model and mutation testing of access control policies. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 667–676. ACM, 2007a. doi: 10.1145/1242572.1242663. URL https://doi.org/10.1145/1242572.1242663.

Evan Martin and Tao Xie. Automated test generation for access control policies via change-impact analysis. In *Third International Workshop on Software Engineering for Secure Systems*, pp. 5. IEEE Computer Society, 2007b. doi: 10.1109/SESS.2007.5. URL https://doi.org/10.1109/SESS.2007.5.

Microsoft Agent Governance Toolkit. Bind every intervention point an adapter evaluates. Git commit e87b74530e1cbae54d5e90bc1b45ecf5c8911b28, 2026. URL https://github.com/microsoft/agent-governance-toolkit/commit/e87b74530e1cbae54d5e90bc1b45ecf5c8911b28.

Seyedmoein Mohsenimofidi, Matthias Galster, Christoph Treude, and Sebastian Baltes. Context engineering for AI agents in open-source software. In *Proceedings of the 23rd International Conference on Mining Software Repositories*, pp. 194–198. Association for Computing Machinery, 2026. doi: 10.1145/3793302.3793350. URL https://doi.org/10.1145/3793302.3793350.

Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. Content moderation for evolving policies using binary question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 561–573. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-industry.54. URL https://aclanthology.org/2023.acl-industry.54/.

Open Policy Agent. Update AGENTS.md guidance for security dependencies. Git commit 6f52916ab1025c26bddc2d5c6d1b19e95e6a4883, 2026. URL https://github.com/open-policy-agent/opa/commit/6f52916ab1025c26bddc2d5c6d1b19e95e6a4883.

Open Policy Agent Gatekeeper Library. Update FlexVolume Drivers PSP to block all when given an empty list. Git commit 102bd7704077cb8f22967b2a75aba2aed4046b02, 2020. URL https://github.com/open-policy-agent/gatekeeper-library/commit/102bd7704077cb8f22967b2a75aba2aed4046b02.

Open Policy Agent Gatekeeper Library. Fix unique service selector behavior when a service selector is missing. Git commit 749ebacafb023738aa95ce468b0503a88d3e2d75, 2021. URL https://github.com/open-policy-agent/gatekeeper-library/commit/749ebacafb023738aa95ce468b0503a88d3e2d75.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392. PMLR, 2025. URL https://proceedings.mlr.press/v267/patil25a.html.

Tzafrir Rehan. Test-driven AI agent definition (TDAD): Compiling tool-using agents from behavioral specifications, 2026. URL https://arxiv.org/abs/2603.08806.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9895–9901. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.779. URL https://aclanthology.org/2021.emnlp-main.779/.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=BsMMc4MEGS.

Sigstore Cosign. Fix handling of policy in verify-attestation. Git commit 3b597b999ad083afd2ce0dba68fa2715ddec0f06, 2022. URL https://github.com/sigstore/cosign/commit/3b597b999ad083afd2ce0dba68fa2715ddec0f06.

SLSA Steering Committee. *Supply-chain Levels for Software Artifacts (SLSA), Version 1.2*. Open Source Security Foundation, 2026. URL https://slsa.dev/spec/v1.2/. Build provenance, artifact verification, and verification summary specifications.

Jan Tretmans. Conformance testing with labelled transition systems: Implementation relations and test generation. *Computer Networks and ISDN Systems*, 29(1):49–79, 1996. doi: 10.1016/S0169-7552(96)00017-7. URL https://doi.org/10.1016/S0169-7552(96)00017-7.

Rachel Tzoref-Brill and Shahar Maoz. Syntactic and semantic differencing for combinatorial models of test designs. In *Proceedings of the 39th International Conference on Software Engineering*, pp. 621–631. IEEE/ACM, 2017. doi: 10.1109/ICSE.2017.63. URL https://doi.org/10.1109/ICSE.2017.63.

WooCommerce Services. Add an AGENTS.md backward-compatibility guardrail. Git commit 2e25e1ac5c16c36c64eeba322dec1010c9ebc8c7, 2026. URL https://github.com/Automattic/woocommerce-services/commit/2e25e1ac5c16c36c64eeba322dec1010c9ebc8c7.

Z.ai. GLM-5.3-Flash: Frontier intelligence, flash cost. Official model release announcement, 2026. URL https://z.ai/blog/glm-5.3-flash. Published 2026-08-26; accessed 2026-08-27.

Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. RuleArena: A benchmark for rule-guided reasoning with LLMs in real-world scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 550–572. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.27. URL https://aclanthology.org/2025.acl-long.27/.

## A Enumerator and Reducer Guarantees

**Minimum-suite hardness.** Take a Set Cover instance with universe $U$ and subsets $S_1, \ldots, S_n$. Create one required rule for each element of $U$ and one candidate context for each $S_i$ whose rule-witness set is $S_i$. Assign every candidate the same required action edge. Any nonempty selection covers that edge, and it covers all required rules exactly when the corresponding subsets cover $U$. A minimum joint suite therefore gives a minimum set cover.

**Exact dynamic program.** Index the $r + k$ required rules and edges and encode coverage as a bit mask. After processing the first $j$ candidates, retain for every reachable mask the shortest subset, breaking ties by context identifier. Excluding candidate $j + 1$ preserves each state; including it unions its coverage mask with each prior mask. Induction on $j$ proves that the final full-mask representative is cardinality-minimum. The implementation matches exhaustive enumeration on 64 deterministic small candidate pools.

**Edge lower bound.** Each pair realizes one directed edge. Covering $k$ edges requires at least $k$ pairs. If the specialized search selects one candidate per edge and also covers every rule, its $k$-pair witness meets the lower bound. This certifies the two-pair Gatekeeper suite and the ten-pair SLSA-inspired suite relative to their declared candidate pools.

# B Complete Action-Level Results

Table 7: Gatekeeper route results by repetition. All columns are end to end. $R$ is raw-strict action; $F$ is exact full decision. Contract action equals $R$ in every row.

| Route | Repetition | $R$ item | $F$ item | $R$ pair | $F$ pair |
|---|---|---|---|---|---|
| Qwen | 1 | 3/24 | 3/24 | 0/12 | 0/12 |
| Qwen | 2 | 2/24 | 2/24 | 0/12 | 0/12 |
| Qwen | 3 | 5/24 | 5/24 | 0/12 | 0/12 |
| | total | 10/72 | 10/72 | 0/36 | 0/36 |
| DeepSeek Flash | 1 | 15/24 | 15/24 | 6/12 | 6/12 |
| DeepSeek Flash | 2 | 19/24 | 18/24 | 7/12 | 6/12 |
| DeepSeek Flash | 3 | 19/24 | 19/24 | 8/12 | 8/12 |
| | total | 53/72 | 52/72 | 21/36 | 20/36 |

Table 8: All output-contract conditions. Conditional scores use valid items or complete pairs; end-to-end scores retain 30-item and 15-pair denominators. "Len." counts request-cap exhaustion.

| Route | Condition | Valid | E2E item/pair | Cond. item/pair | Len. |
|---|---|---|---|---|---|
| GLM | prompt-4096 | 4/30 | 4/30; 0/15 | 4/4; – | 21 |
| GLM | prompt-8192 | 15/30 | 15/30; 6/15 | 15/15; 6/6 | 10 |
| GLM | prompt-12288 | 18/30 | 18/30; 7/15 | 18/18; 7/7 | 6 |
| GLM | schema-4096 | 7/30 | 7/30; 2/15 | 7/7; 2/2 | 17 |
| HY3 | prompt-4096 | 3/30 | 3/30; 0/15 | 3/3; – | 22 |
| HY3 | prompt-8192 | 18/30 | 18/30; 7/15 | 18/18; 7/7 | 6 |
| HY3 | prompt-12288 | 22/30 | 22/30; 9/15 | 22/22; 9/9 | 2 |
| HY3 | schema-4096 | 3/30 | 3/30; 0/15 | 3/3; – | 21 |
| Qwen+ | prompt-4096 | 22/30 | 18/30; 7/15 | 18/22; 7/9 | 0 |
| Qwen+ | prompt-8192 | 20/30 | 17/30; 5/15 | 17/20; 5/7 | 0 |
| Qwen+ | prompt-12288 | 21/30 | 19/30; 7/15 | 19/21; 7/9 | 0 |
| Qwen+ | schema-4096 | 23/30 | 13/30; 2/15 | 13/23; 2/11 | 0 |

Table 9: Non-`ALLOW` full-policy confusion counts pooled over three repetitions. Columns A, H, D, E, and `N` denote predicted actions and terminal route failure.

| Route | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| GLM | `HOLD` | 0 | 15 | 0 | 0 | 3 |
| GLM | `DEMOTE` | 0 | 0 | 12 | 0 | 0 |
| GLM | `ESCALATE` | 0 | 0 | 0 | 17 | 1 |
| HY3 | `HOLD` | 0 | 11 | 0 | 0 | 7 |
| HY3 | `DEMOTE` | 0 | 0 | 12 | 0 | 0 |
| HY3 | `ESCALATE` | 0 | 0 | 0 | 13 | 5 |
| Qwen+ | `HOLD` | 0 | 15 | 0 | 0 | 3 |
| Qwen+ | `DEMOTE` | 0 | 0 | 12 | 0 | 0 |
| Qwen+ | `ESCALATE` | 0 | 0 | 0 | 16 | 2 |

Table 10: GLM output-contract confusion counts. Every condition retains 30 scheduled items.

| Condition | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| prompt-4096 | ALLOW | 3 | 0 | 0 | 0 | 12 |
| | HOLD | 0 | 1 | 0 | 0 | 4 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 0 | 5 |
| prompt-8192 | ALLOW | 8 | 0 | 0 | 0 | 7 |
| | HOLD | 0 | 4 | 0 | 0 | 1 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 3 | 2 |
| prompt-12288 | ALLOW | 10 | 0 | 0 | 0 | 5 |
| | HOLD | 0 | 4 | 0 | 0 | 1 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 4 | 1 |
| schema-4096 | ALLOW | 5 | 0 | 0 | 0 | 10 |
| | HOLD | 0 | 2 | 0 | 0 | 3 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 0 | 5 |

Table 11: HY3 output-contract confusion counts. Every condition retains 30 scheduled items.

| Condition | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| prompt-4096 | ALLOW | 3 | 0 | 0 | 0 | 12 |
| | HOLD | 0 | 0 | 0 | 0 | 5 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 0 | 5 |
| prompt-8192 | ALLOW | 10 | 0 | 0 | 0 | 5 |
| | HOLD | 0 | 2 | 0 | 0 | 3 |
| | DEMOTE | 0 | 0 | 2 | 0 | 3 |
| | ESCALATE | 0 | 0 | 0 | 4 | 1 |
| prompt-12288 | ALLOW | 12 | 0 | 0 | 0 | 3 |
| | HOLD | 0 | 3 | 0 | 0 | 2 |
| | DEMOTE | 0 | 0 | 3 | 0 | 2 |
| | ESCALATE | 0 | 0 | 0 | 4 | 1 |
| schema-4096 | ALLOW | 3 | 0 | 0 | 0 | 12 |
| | HOLD | 0 | 0 | 0 | 0 | 5 |
| | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| | ESCALATE | 0 | 0 | 0 | 0 | 5 |

Table 12: Qwen 3.7 Plus output-contract confusion counts. Every condition retains 30 scheduled items.

| Condition | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| prompt-4096 | ALLOW | 11 | 0 | 0 | 0 | 4 |
| | HOLD | 0 | 2 | 0 | 0 | 3 |
| | DEMOTE | 2 | 1 | 1 | 0 | 1 |
| | ESCALATE | 1 | 0 | 0 | 4 | 0 |
| prompt-8192 | ALLOW | 10 | 0 | 0 | 1 | 4 |
| | HOLD | 0 | 3 | 0 | 0 | 2 |
| | DEMOTE | 1 | 1 | 0 | 0 | 3 |
| | ESCALATE | 0 | 0 | 0 | 4 | 1 |
| prompt-12288 | ALLOW | 10 | 0 | 0 | 2 | 3 |
| | HOLD | 0 | 3 | 0 | 0 | 2 |
| | DEMOTE | 0 | 0 | 2 | 0 | 3 |
| | ESCALATE | 0 | 0 | 0 | 4 | 1 |
| schema-4096 | ALLOW | 10 | 1 | 0 | 0 | 4 |
| | HOLD | 2 | 2 | 0 | 0 | 1 |
| | DEMOTE | 2 | 1 | 0 | 0 | 2 |
| | ESCALATE | 3 | 1 | 0 | 1 | 0 |

Table 13: External policy-revision first-replication confusion matrices. Columns A, H, D, E, and N denote ALLOW, HOLD, DEMOTE, ESCALATE, and terminal route failure; every row retains its full target support.

| Route | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| GLM | ALLOW | 12 | 0 | 0 | 0 | 2 |
| GLM | HOLD | 0 | 1 | 0 | 0 | 2 |
| GLM | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| GLM | ESCALATE | 0 | 0 | 0 | 2 | 0 |
| Luna | ALLOW | 12 | 0 | 0 | 0 | 2 |
| Luna | HOLD | 0 | 1 | 0 | 0 | 2 |
| Luna | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| Luna | ESCALATE | 0 | 0 | 0 | 2 | 0 |
| HY3 | ALLOW | 13 | 0 | 0 | 0 | 1 |
| HY3 | HOLD | 0 | 2 | 0 | 0 | 1 |
| HY3 | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| HY3 | ESCALATE | 0 | 0 | 0 | 2 | 0 |
| Kimi | ALLOW | 13 | 0 | 0 | 0 | 1 |
| Kimi | HOLD | 0 | 2 | 0 | 0 | 1 |
| Kimi | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| Kimi | ESCALATE | 0 | 0 | 0 | 2 | 0 |

| Route | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| MiMo | ALLOW | 13 | 0 | 0 | 1 | 0 |
| MiMo | HOLD | 0 | 2 | 0 | 0 | 1 |
| MiMo | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| MiMo | ESCALATE | 0 | 0 | 0 | 2 | 0 |
| MiniMax | ALLOW | 11 | 0 | 0 | 1 | 2 |
| MiniMax | HOLD | 0 | 1 | 0 | 0 | 2 |
| MiniMax | DEMOTE | 0 | 0 | 1 | 0 | 0 |
| MiniMax | ESCALATE | 0 | 0 | 0 | 2 | 0 |
| OX | ALLOW | 10 | 0 | 0 | 0 | 4 |
| OX | HOLD | 0 | 1 | 0 | 0 | 2 |
| OX | DEMOTE | 0 | 0 | 0 | 0 | 1 |
| OX | ESCALATE | 0 | 0 | 0 | 1 | 1 |
| Qwen+ | ALLOW | 11 | 1 | 0 | 1 | 1 |
| Qwen+ | HOLD | 0 | 2 | 0 | 0 | 1 |
| Qwen+ | DEMOTE | 0 | 0 | 0 | 0 | 1 |
| Qwen+ | ESCALATE | 0 | 0 | 0 | 2 | 0 |

Table 14: End-to-end action recall for every route summarized in the main text. "–" marks actions outside a task's label set.

| Configuration and endpoint | ALLOW | HOLD | DEMOTE | ESCALATE |
|---|---|---|---|---|
| Gatekeeper Qwen, raw-strict | 4/36 | 6/36 | – | – |
| Gatekeeper Flash, raw-strict | 32/36 | 21/36 | – | – |
| Imbalanced Qwen, contract | 143/150 | 110/134 | 0/8 | 5/8 |
| Challenge Qwen, raw-strict | 77/90 | 26/30 | 3/30 | 22/30 |
| Challenge Qwen, contract | 79/90 | 26/30 | 4/30 | 23/30 |
| Challenge Flash, raw-strict | 68/90 | 10/30 | 2/30 | 25/30 |
| Challenge Flash, contract | 68/90 | 9/30 | 0/30 | 13/30 |
| External GLM 5.3, rep. 1 | 12/14 | 1/3 | 1/1 | 2/2 |
| External Luna 5.6, rep. 1 | 12/14 | 1/3 | 1/1 | 2/2 |
| External HY3, rep. 1 | 13/14 | 2/3 | 1/1 | 2/2 |
| External Kimi K2.7, rep. 1 | 13/14 | 2/3 | 1/1 | 2/2 |
| External MiMo 2.5, rep. 1 | 13/14 | 2/3 | 1/1 | 2/2 |
| External MiniMax 2.7, rep. 1 | 11/14 | 1/3 | 1/1 | 2/2 |
| External GLM 5.3 Flash (OX), rep. 1 | 10/14 | 1/3 | 0/1 | 1/2 |
| External Qwen 3.7+, rep. 1 | 11/14 | 2/3 | 0/1 | 2/2 |
| Non-ALLOW GLM, 3 reps. | – | 15/18 | 12/12 | 17/18 |
| Non-ALLOW HY3, 3 reps. | – | 11/18 | 12/12 | 13/18 |
| Non-ALLOW Qwen+, 3 reps. | – | 15/18 | 12/12 | 16/18 |
| Output GLM, prompt-4096 | 3/15 | 1/5 | 0/5 | 0/5 |
| Output GLM, prompt-8192 | 8/15 | 4/5 | 0/5 | 3/5 |
| Output GLM, prompt-12288 | 10/15 | 4/5 | 0/5 | 4/5 |
| Output GLM, schema-4096 | 5/15 | 2/5 | 0/5 | 0/5 |
| Output HY3, prompt-4096 | 3/15 | 0/5 | 0/5 | 0/5 |
| Output HY3, prompt-8192 | 10/15 | 2/5 | 2/5 | 4/5 |
| Output HY3, prompt-12288 | 12/15 | 3/5 | 3/5 | 4/5 |
| Output HY3, schema-4096 | 3/15 | 0/5 | 0/5 | 0/5 |
| Output Qwen+, prompt-4096 | 11/15 | 2/5 | 1/5 | 4/5 |
| Output Qwen+, prompt-8192 | 10/15 | 3/5 | 0/5 | 4/5 |
| Output Qwen+, prompt-12288 | 10/15 | 3/5 | 2/5 | 4/5 |
| Output Qwen+, schema-4096 | 10/15 | 2/5 | 0/5 | 1/5 |
| Factor GLM 5.3, rep. 1 | 7/15 | 1/5 | 0/5 | 1/5 |
| Factor Luna 5.6, rep. 1 | 14/15 | 4/5 | 3/5 | 5/5 |
| Factor HY3, rep. 1 | 3/15 | 0/5 | 0/5 | 0/5 |
| Factor Kimi K2.7, rep. 1 | 10/15 | 3/5 | 0/5 | 2/5 |
| Factor MiMo 2.5, rep. 1 | 9/15 | 0/5 | 0/5 | 2/5 |
| Factor MiniMax 2.7, rep. 1 | 6/15 | 0/5 | 0/5 | 1/5 |
| Factor Qwen 3.7+, rep. 1 | 14/15 | 5/5 | 2/5 | 4/5 |
| SLSA Qwen, raw-strict | 60/60 | 20/20 | 20/20 | 20/20 |
| SLSA Qwen, contract | 60/60 | 20/20 | 1/20 | 7/20 |
| SLSA Flash, raw-strict | 59/60 | 17/20 | 17/20 | 17/20 |
| SLSA Flash, contract | 59/60 | 17/20 | 15/20 | 17/20 |
| Graph Mistral, contract | 9/36 | 18/24 | 0/24 | 13/36 |
| Graph Phi-4, contract | 21/36 | 5/24 | 0/24 | 1/36 |

Table 15: Factor-complete first-replication confusion matrices. N is a terminal route failure. Every row retains its full target support.

| Route | Target | ALLOW | HOLD | DEMOTE | ESCALATE | N |
|---|---|---|---|---|---|---|
| GLM | ALLOW | 7 | 0 | 0 | 0 | 8 |
| GLM | HOLD | 0 | 1 | 0 | 0 | 4 |
| GLM | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| GLM | ESCALATE | 0 | 0 | 0 | 1 | 4 |
| Luna | ALLOW | 14 | 1 | 0 | 0 | 0 |
| Luna | HOLD | 0 | 4 | 0 | 0 | 1 |
| Luna | DEMOTE | 2 | 0 | 3 | 0 | 0 |
| Luna | ESCALATE | 0 | 0 | 0 | 5 | 0 |
| HY3 | ALLOW | 3 | 0 | 0 | 0 | 12 |
| HY3 | HOLD | 0 | 0 | 0 | 0 | 5 |
| HY3 | DEMOTE | 0 | 0 | 0 | 0 | 5 |
| HY3 | ESCALATE | 0 | 0 | 0 | 0 | 5 |
| Kimi | ALLOW | 10 | 0 | 0 | 0 | 5 |
| Kimi | HOLD | 0 | 3 | 0 | 0 | 2 |
| Kimi | DEMOTE | 1 | 0 | 0 | 0 | 4 |
| Kimi | ESCALATE | 0 | 0 | 0 | 2 | 3 |
| MiMo | ALLOW | 9 | 0 | 0 | 2 | 4 |
| MiMo | HOLD | 0 | 0 | 0 | 0 | 5 |
| MiMo | DEMOTE | 3 | 0 | 0 | 0 | 2 |
| MiMo | ESCALATE | 0 | 0 | 0 | 2 | 3 |
| MiniMax | ALLOW | 6 | 0 | 0 | 0 | 9 |
| MiniMax | HOLD | 0 | 0 | 0 | 0 | 5 |
| MiniMax | DEMOTE | 2 | 0 | 0 | 0 | 3 |
| MiniMax | ESCALATE | 0 | 0 | 0 | 1 | 4 |
| Qwen+ | ALLOW | 14 | 0 | 0 | 1 | 0 |
| Qwen+ | HOLD | 0 | 5 | 0 | 0 | 0 |
| Qwen+ | DEMOTE | 1 | 0 | 2 | 0 | 2 |
| Qwen+ | ESCALATE | 0 | 0 | 0 | 4 | 1 |

Table 16: Gatekeeper aggregate raw-strict confusion counts over three repetitions (left) and SLSA-inspired transition-graph contract-action confusion counts (right). Columns A, H, D, E, and N denote ALLOW, HOLD, DEMOTE, ESCALATE, and terminal route failure.

**Gatekeeper**

| Route | Target | A | H | N |
|---|---|---|---|---|
| Qwen | ALLOW | 4 | 0 | 32 |
| Qwen | HOLD | 0 | 6 | 30 |
| Flash | ALLOW | 32 | 0 | 4 |
| Flash | HOLD | 0 | 21 | 15 |

**SLSA-inspired graph**

| Route | Target | A | H | D | E | N |
|---|---|---|---|---|---|---|
| Mistral | ALLOW | 9 | 8 | 0 | 9 | 10 |
| Mistral | HOLD | 2 | 18 | 0 | 4 | 0 |
| Mistral | DEMOTE | 3 | 0 | 0 | 0 | 21 |
| Mistral | ESCALATE | 2 | 3 | 0 | 13 | 18 |
| Phi-4 | ALLOW | 21 | 0 | 0 | 0 | 15 |
| Phi-4 | HOLD | 1 | 5 | 0 | 0 | 18 |
| Phi-4 | DEMOTE | 3 | 12 | 0 | 0 | 9 |
| Phi-4 | ESCALATE | 0 | 0 | 0 | 1 | 35 |

Table 17: Imbalanced-suite and balanced-challenge full-contract confusion counts. Columns are terminal predicted actions; N is a contract failure.

| Configuration | Target | ALLOW | HOLD | DEMOTE | ESCALATE | N |
|---|---|---|---|---|---|---|
| Imbalanced Qwen | ALLOW | 143 | 0 | 0 | 0 | 7 |
| Imbalanced Qwen | HOLD | 4 | 110 | 0 | 2 | 18 |
| Imbalanced Qwen | DEMOTE | 8 | 0 | 0 | 0 | 0 |
| Imbalanced Qwen | ESCALATE | 0 | 0 | 0 | 5 | 3 |
| Challenge Qwen | ALLOW | 79 | 0 | 0 | 0 | 11 |
| Challenge Qwen | HOLD | 0 | 26 | 0 | 0 | 4 |
| Challenge Qwen | DEMOTE | 8 | 4 | 4 | 4 | 10 |
| Challenge Qwen | ESCALATE | 0 | 0 | 0 | 23 | 7 |
| Challenge Flash | ALLOW | 68 | 4 | 0 | 8 | 10 |
| Challenge Flash | HOLD | 13 | 9 | 0 | 0 | 8 |
| Challenge Flash | DEMOTE | 25 | 3 | 0 | 0 | 2 |
| Challenge Flash | ESCALATE | 0 | 3 | 0 | 13 | 14 |

Table 18: Balanced-challenge raw-strict action confusion counts. N marks a terminal extraction failure.

| Configuration | Target | ALLOW | HOLD | DEMOTE | ESCALATE | N |
|---|---|---|---|---|---|---|
| Challenge Qwen | ALLOW | 77 | 0 | 0 | 0 | 13 |
| Challenge Qwen | HOLD | 0 | 26 | 1 | 0 | 3 |
| Challenge Qwen | DEMOTE | 7 | 3 | 3 | 4 | 13 |
| Challenge Qwen | ESCALATE | 0 | 0 | 0 | 22 | 8 |
| Challenge Flash | ALLOW | 68 | 4 | 5 | 13 | 0 |
| Challenge Flash | HOLD | 13 | 10 | 7 | 0 | 0 |
| Challenge Flash | DEMOTE | 25 | 3 | 2 | 0 | 0 |
| Challenge Flash | ESCALATE | 0 | 3 | 2 | 25 | 0 |

Table 19: SLSA-inspired raw-strict and contract-action confusion counts. N is the endpoint's terminal failure category.

| Route | Endpoint | Target | ALLOW | HOLD | DEMOTE | ESCALATE | N |
|---|---|---|---|---|---|---|---|
| Qwen | raw-strict | ALLOW | 60 | 0 | 0 | 0 | 0 |
| Qwen | raw-strict | HOLD | 0 | 20 | 0 | 0 | 0 |
| Qwen | raw-strict | DEMOTE | 0 | 0 | 20 | 0 | 0 |
| Qwen | raw-strict | ESCALATE | 0 | 0 | 0 | 20 | 0 |
| Qwen | contract | ALLOW | 60 | 0 | 0 | 0 | 0 |
| Qwen | contract | HOLD | 0 | 20 | 0 | 0 | 0 |
| Qwen | contract | DEMOTE | 0 | 0 | 1 | 0 | 19 |
| Qwen | contract | ESCALATE | 0 | 0 | 0 | 7 | 13 |
| Flash | raw-strict | ALLOW | 59 | 0 | 0 | 0 | 1 |
| Flash | raw-strict | HOLD | 0 | 17 | 0 | 0 | 3 |
| Flash | raw-strict | DEMOTE | 0 | 0 | 17 | 0 | 3 |
| Flash | raw-strict | ESCALATE | 0 | 0 | 0 | 17 | 3 |
| Flash | contract | ALLOW | 59 | 0 | 0 | 0 | 1 |
| Flash | contract | HOLD | 0 | 17 | 0 | 0 | 3 |
| Flash | contract | DEMOTE | 0 | 0 | 15 | 0 | 5 |
| Flash | contract | ESCALATE | 0 | 0 | 0 | 17 | 3 |