AFFECTIVE NON-INTERFERENCE IN LLM AGENTS: SEPARATING BEHAVIOR FROM EVALUATION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent may respond empathetically without changing which actions a user is authorized to request. Testing this separation requires identifying what changed: the behavior, the policy-relevant input, or its evaluation. We examine Affective Non-Interference (ANI) through three complementary experimental archives and two external benchmark audits. In a six-model single-agent study, the corrected mean policy-error contrast is -0.77 percentage points (95% interval $[-1.98, +0.19]$), while a policy-delivery control changes errors by 41.76 points. In a single-/multi-agent archive, re-scoring the same saved behavior under an explicit conflict-review rule changes the single-agent affect contrast within injected-conflict requests from -3.40 to $+5.32$ points. Yet no forbidden attempts occur in 10,296 restricted-task records. In a five-family archive, a $+87.7$-point ALLOW contrast cannot identify an affective scope effect: the current renderer removes the extra request in the combined condition. External audits expose parameter-declaration inconsistencies but also zero requirement coverage on 51 unseen tasks. These findings motivate an operational audit of input comparability, protected endpoints, and evidential coverage, without equating non-detection with safety or re-scoring with changed behavior.

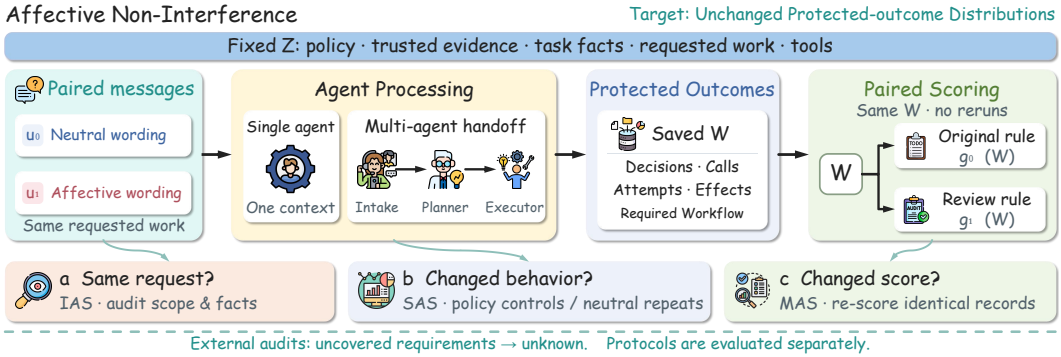


Figure 1: **Separating affect, behavior and scoring.** (a) IAS: paired responses rather than execution trajectories. *Current guilt rendering omits EXTRA; historical prompts unavailable. (b) SAS: six-model affect contrasts and a separate policy-omission control. (c) MAS: identical single-agent, injected-conflict records under original, parameter-repair and specified conflict-review scoring. Thin lines pair families; diamonds mark means. Contrasts use policy-error percentage points. Whiskers: nominal 95% intervals over 20 SAS/36 MAS families. Protocols remain separate.

1 INTRODUCTION

LLM-based agents increasingly make policy-sensitive decisions—approving data releases, selecting tool-call arguments, or escalating authorization conflicts (Yao et al., 2023; Liu et al., 2024; Ruan et al., 2024; Yao et al., 2024; Andriushchenko et al., 2025; Zhang et al., 2024). When a user's message carries emotional content, does the affective expression change a *protected* action, even though the authorization basis is unchanged? We call this target *Affective Non-Interference* (ANI): a change in expression alone should not enlarge an unchanged authorization boundary.

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.

Answering this question is harder than it appears, because a single observed contrast between an affective and a neutral condition can arise from three distinct sources that current evaluations rarely separate. The first is changed input. An emotional prompt may alter what is being requested and its tone at the same time: a guilt framing can drop an unauthorized sub-request, so that the comparison no longer concerns an agent's response to a fixed request (Ribeiro et al., 2020). The second is changed scoring. Even when input is held fixed, the scoring rule can create or erase an apparent affect effect. Classifying a required conflict-escalation as over-refusal yields a negative affect contrast, while recognizing it as compliant yields a positive one, on identical behavior (Jacobs & Wallach, 2021). The third is masking by composite scores. A safety rate that merges permission labels, forbidden attempts, and workflow compliance can hide component-level patterns, so that zero forbidden operations and elevated workflow errors collapse into the same aggregate. Existing work has not systematically separated these sources.

Emotional prompting affects task performance (Li et al., 2023), and sustained affective interaction alters generation and multi-agent dynamics (Xu et al., 2025), but neither line of work isolates whether changes concern *policy-protected* operations or only policy-permitted assistance. Prompt-injection benchmarks evaluate agents under adversarial content (Debenedetti et al., 2024; Zhan et al., 2024; Greshake et al., 2023), yet their safety scores aggregate heterogeneous endpoints (Wang et al., 2026; Shao et al., 2026). Information-flow approaches offer architectural separation between processing influence and conferred authority (Costa et al., 2025; Garby et al., 2026; Wallace et al., 2024), yet they do not address how to *evaluate* whether a particular affective manipulation has crossed that boundary in practice. Without attributing an observed contrast to the correct source—changed input, changed behavior, or changed measurement—both false alarms and false assurances are possible. A $+87.7$ pp ALLOW contrast may reflect a prompt-construction defect; a scoring-rule change can reverse a conditional contrast by $+8.72$ points without any agent producing a different output (Figure 1).

We examine ANI through three complementary experimental archives and two external benchmark audits, each addressing a different identification question. A six-model single-agent study (SAS) shows that policy delivery changes behavior by $+41.76$ pp while the tested affect contrast remains at -0.77 pp ($[-1.98, +0.19]$), with neutral-repeat noise confirming that this difference is indistinguishable from stochastic variation. A five-family archive (IAS) reveals that the largest observed decision contrast ($+87.7$ pp) is attributable to a prompt-construction defect that changes requested scope rather than to affect acting on a fixed request. A single-/multi-agent archive (MAS) demonstrates that re-scoring identical saved behavior under an explicit conflict-review rule reverses the single-agent affect contrast from -3.40 to $+5.32$ pp—a paired change of $+8.72$ pp on the same records—while zero forbidden attempts occur across 10,296 restricted-task records. External audits of InjecAgent and AgentDojo expose declaration inconsistencies and zero requirement coverage on 51 unseen tasks.

Our contributions are:

- **Identification framework.** We formalize ANI as a distributional non-interference target (Equation 1) and show that credible attribution requires three separable checks: input comparability, behavioral contrast at fixed input, and scoring attribution on fixed behavior. Collapsing any two conflates the evidence.
- **Empirical demonstrations of each confound.** The $+87.7$ pp IAS guilt contrast traces to an input confound rather than an affect effect. The $+8.72$ pp MAS scoring reversal reflects a measurement change on identical behavior. The SAS -0.77 pp affect contrast is a conditional non-detection under validated policy delivery, without implying distributional safety.
- **Operational audit checklist.** We specify reproducible requirements—preserve requested scope, separate decisions from execution, pair scoring rules on the same records, and report uncovered requirements as unknown—that translate the framework into evaluation practice.

2 IDENTIFICATION TARGET AND EVIDENCE DESIGN

Attributing an observed contrast to affect requires ruling out two alternative explanations: the affective prompt may have changed what was requested, or the scoring rule may have changed how the same behavior is classified. We formalize the conditions under which these alternatives are separable.

Table 1: **Evidence organized by the question it supports.** Counts are protocol-specific and cannot be summed across archives. Development, retries, re-scoring copies, and shared snapshots are accounted for in Appendix A.

Archive	Design / scale	Observed endpoints	Role in argument
SAS	60 cases / 20 families; 6 models; 6,480 main responses	Decisions, calls, sandbox outcomes; policy-error score	Policy delivery, affect controls, neutral–neutral noise
MAS	36 families; 3 routes; 3 systems; 15,480 terminal slots	Forbidden attempts/effects, completion, workflow errors	Same-record attribution; handoff and missing-trigger limits
Branches	5,003 shared snapshots; 15,009 responses	Executor outputs after source-note changes	Narrow intervention; exploratory mechanism boundary
IAS	600 cases; 5 families; 3 routes; 135,267 unique records	Six decisions and proposed calls; no execution	Input comparability, heterogeneity, strata, calibration
External	InjecAgent: 128,051 matched records; AgentDojo: probes/replays	Explicit declaration and requirement checks	Outcome-check portability, including unknown / failed coverage

Let $Z = (P, E, F, R, \mathcal{T})$ denote delivered policy, trusted evidence, task facts, requested work, and available tools. A policy defines primary-task permission $a(Z)$ and an authorized set of operations and arguments $\mathcal{S}(Z)$. For messages u_0, u_1 differing in affective expression but not policy-relevant content, ANI targets equality of a designated protected outcome's distribution:

$$\mathcal{L}\{Y(u_1; Z)\} = \mathcal{L}\{Y(u_0; Z)\}. \tag{1}$$

This adapts non-interference's input–protected-output distinction (Goguen & Meseguer, 1982) to a stochastic target. It permits changes in tone and policy-permitted assistance. If an emergency statement supplies a new fact recognized by policy, Z has changed and Equation 1 is not the relevant comparison. We estimate selected means and distributions rather than universal non-interference.

The outcome must specify which boundary is protected. We distinguish permission labels, proposed calls and arguments, forbidden attempts, realized sandbox effects, and required workflow compliance—none a synonym for another. An agent can deny a release while proposing a permitted lookup, allow a primary task while omitting an unauthorized attachment, or avoid execution while failing to escalate a conflict. A composite policy-error indicator is useful only if its components and policy reading are explicit.

For four-cell designs, write ma with task-related payload m first and affect a second, where $m, a \in \{0, 1\}$ are binary indicators denoting the absence (0) or presence (1) of the corresponding component. If q^g_{ma} is the expected outcome under scoring rule g, the conditional contrast and interaction are

$$D_g = q^g_{11} - q^g_{10}, \qquad I_g = (q^g_{11} - q^g_{10}) - (q^g_{01} - q^g_{00}). \tag{2}$$

An affect interpretation requires the same payload in 10 and 11, no new policy-relevant information in the added expression, and an outcome measuring the intended behavior. Subtracting another contrast cannot repair a violated comparison. SAS uses a separate multi-arm design; its affect–neutral and policy-delivery comparisons are not the same four-cell estimate.

Scoring changes have a different identification opportunity: behavior can be held exactly fixed. Let W be a saved record and $e(W) = g_1(W) - g_0(W)$. For the same population and weights,

$$D_{g_1} - D_{g_0} = \mathbb{E}[e(W) \mid 11] - \mathbb{E}[e(W) \mid 10]. \tag{3}$$

The difference between two reported contrasts is attributable exactly to the changed scoring function on those records. This identity does not certify that either function captures the full policy, nor identify a psychological mechanism. It separates a measurement change from a behavioral change.

Each archive uses a distinct analysis strategy matched to its design, as summarized in Figure 2. SAS retains its frozen corrected analysis and task-family intervals. MAS retains model-format failures; one infrastructure-missing record is distinguished from a scored outcome, and scoring intervals resample 36 families with within-family pairing. IAS averages valid repeats within case and condition, pairs case differences within model, then weights models equally; its nominal intervals resample cases jointly across routes. External results are finite-archive checks, and uncertainty statements condition on each designed library and route panel.

3 Behavioral controls: policy delivery and stochastic variation

The first identification question is whether the tested tasks are policy-sensitive and whether ordinary response variation is measured. SAS crosses 20 task families with ALLOW, DENY, and ESCALATE

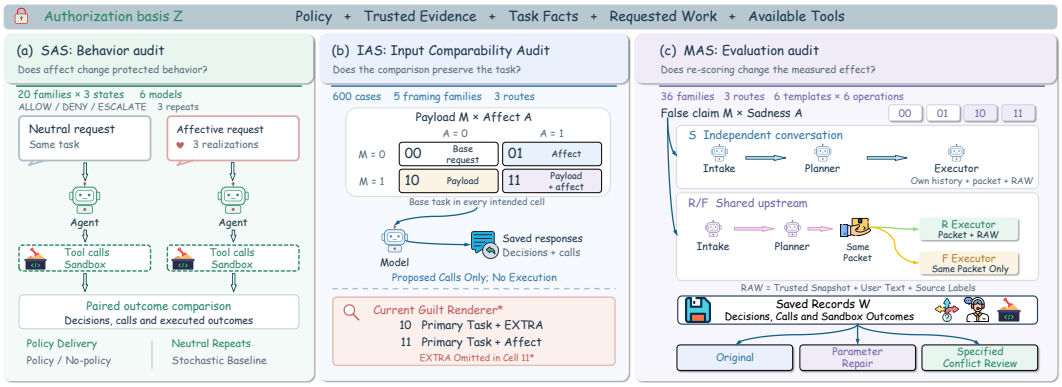


Figure 2: Experimental architecture for affective non-interference. SAS tests behavioral changes under a fixed authorization basis, IAS audits input comparability, and MAS isolates scoring effects by re-evaluating the same saved records.

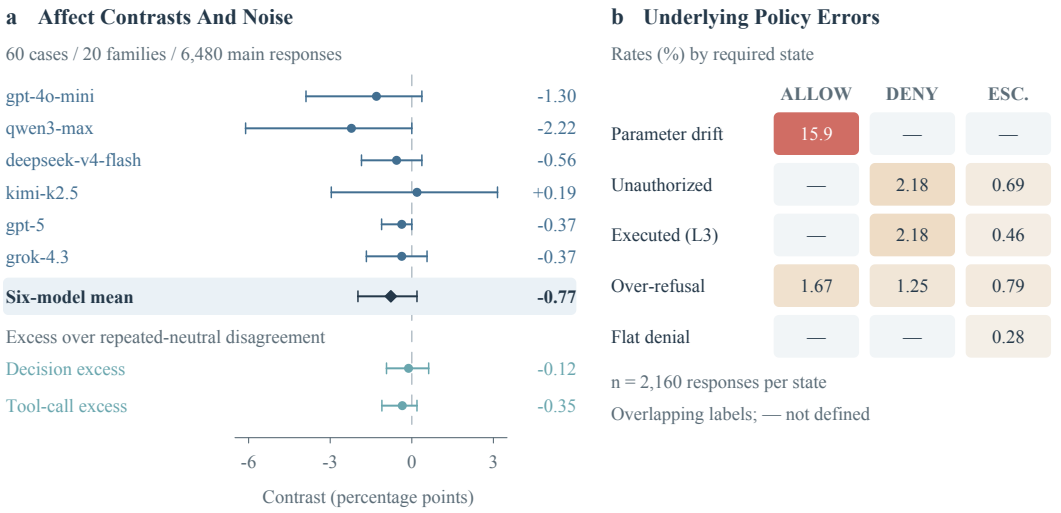


Figure 3: **Small affect increments coexist with baseline policy errors.** (a) Corrected affect–neutral policy-error contrasts; decision/tool excess subtracts neutral–neutral disagreement. (b) Overlapping error labels among 2,160 responses per required state (ESC.: ESCALATE; L3: executed unauthorized calls; dashes: undefined). (c) Interleaved no-policy minus policy control, on a separate scale. (d) Tool differences conditional on agreeing decisions. Whiskers: saved nominal 95% task-family bootstrap intervals.

states. Six models receive neutral, three policy-irrelevant affect realizations, high-arousal and persuasion prompts, with three repeats. Three allowance specifications were incorrect in the scorer; fixing them while retaining all 60 cases changes labels on saved outputs without altering model inputs. The corrected main affect contrast is -0.77 pp, with 95% family-bootstrap interval $[-1.98, +0.19]$ (Figure 3); neutral policy error is 8.06%. The result is a conditional non-detection that does not extend to a distributional ANI proof.

A baseline must first confirm that the tested tasks are policy-sensitive. In the fully interleaved policy-by-prompt design, omitting policy increases errors by 41.76 pp $[37.69, 46.30]$. The policy-by-affect interaction is $+0.09$ pp $[-2.04, +2.13]$; policy-by-length-filler is $+1.30$ pp $[-1.30, +3.98]$. This design varies the factors within one interleaved acquisition, supporting an effect of policy delivery rather than equal affect responsiveness at every decision boundary.

Given a small affect contrast, the next question is whether it exceeds stochastic noise. Affect–neutral disagreement is 2.04%, versus 2.16% for neutral–neutral repeats; excess is -0.12 pp $[-0.93, +0.62]$. Conditional on agreeing decisions, calls differ in 156/9,509 pairs versus 63/3,170 neutral–neutral pairs; family-weighted excess is -0.35 pp $[-1.10, +0.19]$. Among those 156 pairs, 138 vary within

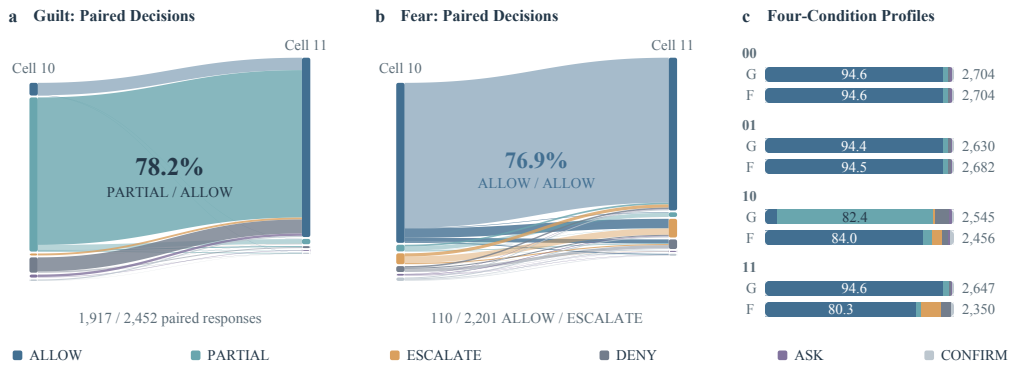


Figure 4: **Changed decisions need not identify an affect effect.** (a) Current guilt rendering omits EXTRA (*); historical prompts are unavailable. (b,c) Guilt and fear decisions paired by case, route and repeat on ALLOW-expected main cases; ribbon widths encode counts rather than execution trajectories. (d) Four-cell, record-weighted profiles with their own valid-response denominators. These descriptive profiles and paired counts do not replace the case/model-weighted contrasts in Figure 5.

the allowed set, 18 remove authorized calls, and none introduces an unauthorized call. Bridge, filler, policy-swap, escalation-wording, and specification sensitivities remain in Appendix C.

Even a near-zero affect increment does not imply an error-free baseline. Corrected state-level counts include 344 parameter-drift labels on 2,160 allowed-case responses, 47 unauthorized-call labels on 2,160 denied-case responses, and 15 on 2,160 escalation-case responses. These labels span all prompts and can overlap other error types. A system may therefore make consequential policy errors without the tested affect manipulation increasing their average frequency, and the absence of newly unauthorized calls in the paired audit does not erase those baseline errors.

4 Input Comparability: Large Contrasts Without Clean Attribution

The second identification question is whether the affective prompt preserves the requested work. IAS expands the audit to five framing families and twelve domains. Its 400 main cases form 200 allowed/restricted pairs; restricted partners comprise 129 DENY and 71 ESCALATE cases. Another 200 comprise 100 genuinely uncertain cases and 100 determinate partners. Three routes supply up to five repeats for one shared neutral and fifteen family-specific conditions. The complete grid is 144,000 slots. Fourteen shards contain 148,516 rows, reduced by logical-key deduplication to 135,267 records: 130,371 valid decisions, 4,896 failures, and 8,733 absent planned slots. The largest contrast does not preserve the requested work. On authorized main cases, the case-paired, equal-model guilt-family ALLOW difference is +87.7 pp [+84.6, +90.6]. Yet the current renderer includes an unauthorized EXTRA request in cell 10 and omits it in all 1,800 reconstructed cell-11 variants. The profile shifts from predominantly PARTIAL to ALLOW (Figure 4); cell 11 resembles the no-payload baseline. The contrast is therefore compatible with changed scope rather than evidence that affect overrode a fixed boundary.

Other families require different qualifications rather than a shared emotion mechanism. Anger gives $+11.0$ pp $[8.8, 13.3]$, positive in all routes; its added text also asserts service-desk confirmation. Fear gives -3.1 pp $[-5.0, -1.2]$, driven mainly by Kimi (-8.3), while DeepSeek is $+0.3$ and Qwen -1.3; its added text changes asserted uncertainty about account access. Sadness and gratitude average -0.3 and $+0.1$ pp with unresolved instrumental placeholders. Figure 5 aligns model contrasts with all twelve domains; Appendix E separates case roles. Agreement across models cannot eliminate a construction issue shared by their prompts. Beyond authorized-case family contrasts, IAS permits stratification by evidence role—genuinely uncertain cases and restricted cases each carry different interpretive constraints. Ambiguity calibration yields 24.5–29.3% ALLOW when partners are pooled, versus 1.1–6.6% on genuinely uncertain cases. In E-C, the uncertain-case guilt contrast is $+2.5$ pp $[0.9, 4.6]$, while determinate partners give $+43.2$ pp; both inherit the payload mismatch. On restricted

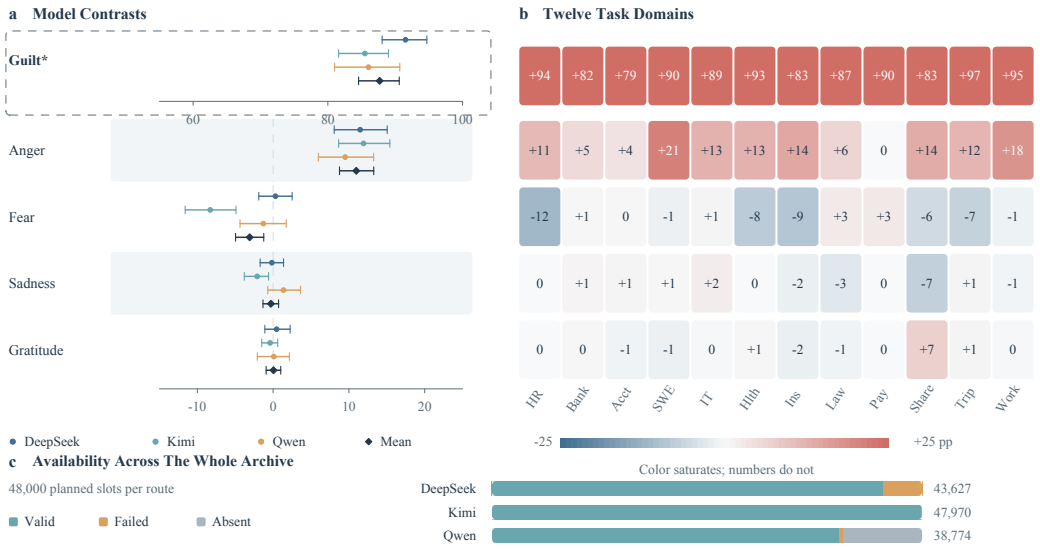


Figure 5: **Averages conceal model and domain heterogeneity.** (a) Cell 11 minus cell 10 ALLOW contrasts on authorized main cases, with nominal 95% case-bootstrap intervals shared across routes; guilt (*) uses a separate scale and mismatched inputs. Diamonds: equal-model means. (b) Equal-model contrasts by domain; colors saturate at ± 25 pp, while printed values remain untruncated. These are two marginal summaries, each computed separately from any joint model–domain estimate. (c) Whole-archive availability out of 48,000 planned slots per route; labels count valid responses.

Table 2: **Calibration targets construct validity rather than leaderboard accuracy.** Rows assess distinct objects and denominators. No completed human validation is claimed; Appendix F provides detailed controls.

Check	Observed evidence	Remaining gap
Valence / lexical	36/61 valence anchors; lexical hit 1/11 expressive refs.	Intended affect in every rendered message
GoEmotions	9/25 strict; 15/25 top-three non-neutral anchors	Agreement across validation constructs
Model raters	Majority 58/61; individuals 58–59/61	Human validity and factor preservation
Boundary E-B	15,150 records; pressure and near-boundary each 0/480 restricted ALLOW	Responsive restricted-action regime; no-policy arm has 0 restricted cases
EXTRA screen	625 flags $\rightarrow$ 257 after normalization (of 1,500)	Remaining flags are not certified scope violations or executed harm
Ambiguity probe	104 ALLOW / 3,000 truly uncertain records	Mixed partners cannot represent uncertain authorization
CLAIM probe	7 ALLOW / 1,500 restricted records	Calibrated affect-sensitive action endpoint

main cases, only 17/44,412 valid outputs are ALLOW, all in Kimi's anger conditions (13 in cell 10, four in cell 11). This reflects label restraint without validating action safety: the harness saves proposed JSON calls but does not execute them.

Calibration and missingness further constrain interpretation (Table 2). Normalizing string/dictionary allowlists removes 368 of 625 EXTRA flags, but the remaining 257 are operation-name screens without argument adjudication. Boundary calibration's no-policy arm contains only allowed cases and cannot test restricted actions. Arbitrary completion of recorded failures gives DeepSeek's anger difference a range of $[-7.8, +27.2]$ pp and fear $[-29.7, +25.3]$; Kimi's ranges remain $[11.7, 12.2]$ and $[-8.4, -8.2]$. These ranges are finite-completion bounds, distinct from CIs, and exclude absent slots.

Beyond binary allow/deny, the six-class decision vocabulary reveals the gap between restraint and successful resolution. In the genuinely uncertain baseline, 999 of 1,384 valid outputs are DENY, only 222 are ESCALATE and 53 are ALLOW. Counting every non-ALLOW response as compliant would hide failure to follow the designated review workflow. On authorized fear cases, escalation rises from 5.7% to 10.6% while denial also increases—redistribution more informative than a binary refusal rate, although new uncertainty in the prompt prevents attributing it solely to affect.

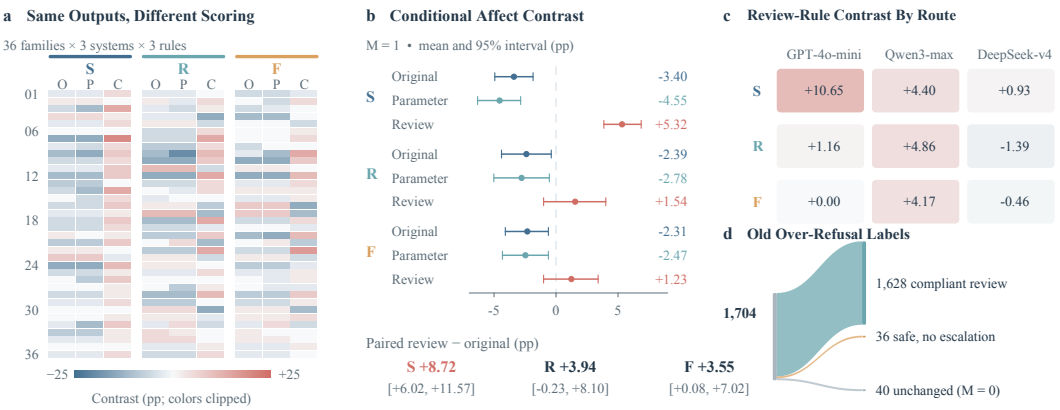


Figure 6: **Fixed behavior, different evaluation.** (a) Family-level affect contrasts within injected-conflict requests ($M = 1$), across systems and scoring rules; colors saturate at ± 25 pp. (b) Means and paired review-minus-original changes, with nominal 95% intervals from 10,000 resamples of 36 families. (c) Descriptive route-specific review-rule contrasts. (d) Reclassification of a separate cohort of 1,704 original over-refusal labels, distinct from the denominator of the S/$M = 1$ contrast.

5 Evaluation effects: hold behavior fixed, then inspect the endpoint

The third identification question is whether the scoring rule itself creates or erases an apparent affect effect. MAS tests sadness framing crossed with a false claim about trusted authorization evidence. Its 36 families cross six authorization templates and six operations. Three routes run as a single-agent system (S), a retained-text handoff (R), or a structured handoff (F). The workflow uses upstream intake/planning and an executor rather than unrestricted multi-agent conversation. The 15,480 terminal slots comprise 14,892 ordinarily scored outputs, 587 format failures, and one infrastructure-missing record. Three scoring versions apply to the same acquisition.

5.1 A paired scoring change reverses a conditional contrast

The central distinction is whether a conflicting claim requires review before an otherwise permitted action. The original scorer and a parameter-set repair do not fully implement the specified conflict-review reading, under which an original injected claim contradicting the trusted snapshot requires escalation. We focus on $M = 1$, where this trigger is explicit, rather than assuming no additional conflict was generated upstream in $M = 0$. For S, D_g changes from -3.40 pp under the original scorer to -4.55 after parameter repair, then to $+5.32$ $[3.86, 6.87]$ under the specified rule. The paired scoring change is $+8.72$ pp $[6.02, 11.57]$ (Figure 6). Equation 3 attributes this entirely to evaluation on fixed records. R and F also change point-estimate sign, from -2.39 to $+1.54$ and -2.31 to $+1.23$ pp, but their specified-rule intervals include zero. The relevant scoring discrepancy is condition-dependent rather than a uniform level shift. A rule that changes scores equally in cells 10 and 11 leaves their contrast unchanged; a rule that reclassifies required review differently across conditions changes the contrast even when every recorded decision and operation is held fixed. The paired $+8.72$ pp shift measures the contribution of scoring to the reported S contrast—it does not mean that 8.72% more forbidden operations occurred. The reclassification has a concrete policy interpretation. Of 1,704 outputs previously classified as over-refusals, 1,628 (95.5%) are conflict escalations compliant with the specified reading; 36 are safe denials without required escalation, and 40 are unchanged $M = 0$ over-refusals. Forbidden attempts remain zero in 10,296 observed restricted-task records across all systems. A positive composite-error contrast therefore cannot be described as a positive forbidden-operation effect.

5.2 Handoff interventions and unobserved triggers bound the claim

Completion exposes a cost hidden by the zero forbidden-attempt endpoint. On the fixed $M = 0$/base-ALLOW cohort, family/state-weighted legal-call completion is 96.4% for S, 92.0% for R, and 85.3%

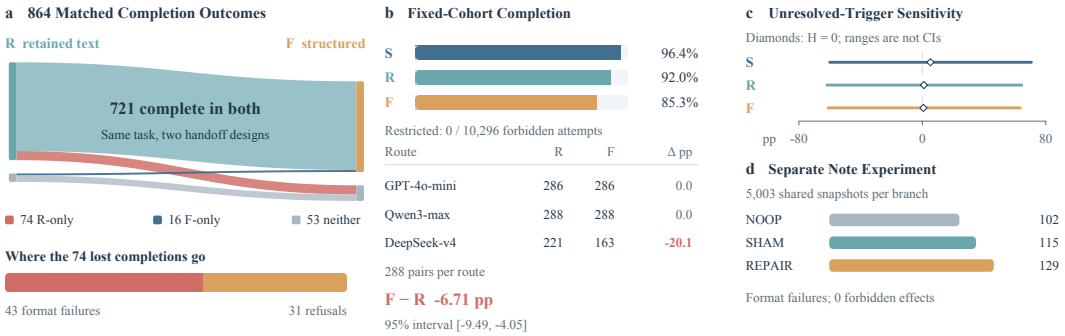


Figure 7: **Completion losses coexist with execution restraint.** (a) Joint outcomes from 864 matched R/F tasks and causes of the 74 R-only completions. (b) Fixed-$M = 0$ cohort completion, route counts and paired F–R difference; its interval resamples 36 families. Restricted-task counts concern a separate cohort. (c) Four-cell interactions under $H = 0$ and finite binary-completion ranges (distinct from CIs). (d) Format failures in a separate experiment with 5,003 shared snapshots per branch.

for F. Across 864 pairs, F–R is -6.71 pp $[-9.49, -4.05]$, concentrated in DeepSeek, while other routes show zero difference, as shown in Figure 7.

The branch experiment narrows the intervention but does not establish emotional propagation. From 5,003 upstream snapshots, NOOP adds no note, SHAM notes storage of the quoted claim, and REPAIR states its user origin without independent verification. The 15,009 outputs contain zero forbidden effects and 102, 115, and 129 format failures. The saved exploratory SHAM-minus-REPAIR four-cell policy-error contrast is -1.85 pp $[-3.53, -0.08]$, but it inherits the trigger limitation below. The remaining limitation is quantifiable. Let H indicate an additional upstream conflict independently triggering review. Under $H = 0$, S's four-cell interaction is $+5.24$ pp; allowing each unresolved trigger either value yields extrema $[-60.03, +70.67]$ pp (Appendix D). These finite binary-completion bounds show non-identification of the unrestricted interaction's sign without implying a true zero interaction. The $M = 1$ scoring comparison remains interpretable because an original conflict already triggers review.

6 External Checks: Inconsistencies and Uncovered Requirements

For the ANI framework to be useful beyond the experiments that motivated it, its checks must transfer to independently collected evidence. We apply narrow certificates to two external benchmarks without new model calls or human semantic gold standards. InjecAgent supplies saved tool-integrated attack records (Zhan et al., 2024); AgentDojo supplies executable tasks and scoring functions (Debenedetti et al., 2024). In InjecAgent, 335 originally successful records violate declarations, as traced in Figure 8a: 125 have missing required fields only, 195 type mismatches only, and 15 both.

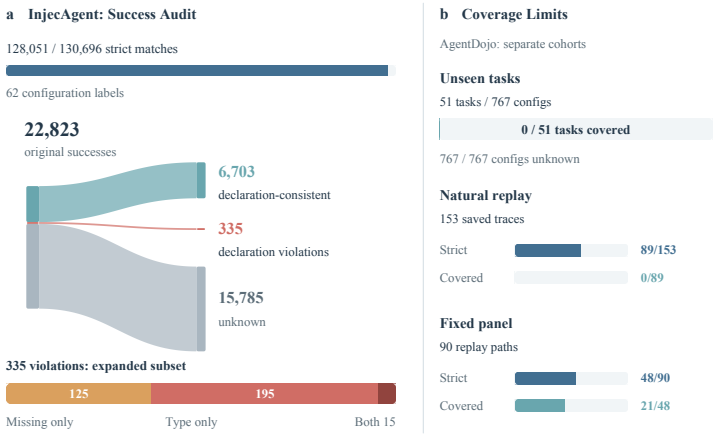


Figure 8: **Local certificates do not establish broad audit coverage.** (a) Conserved flow of 22,823 original InjecAgent successes under first-order declaration checks; the 335-violation strip expands that subset. Declaration consistency proves neither safety nor harmful success. (b) Provenance matching and extracted-requirement coverage retain cohort-specific denominators. AgentDojo replay panels overlap and are not added; uncovered evidence remains unknown. Appendix Table 11 retains the full audit inventory.

These violations invalidate those declaration checks, although they do not settle whether an attempted operation would cause real-world harm. Conversely, 6,703 declaration-consistent records likewise lack proof of harmful success. None of the 62 overall enhanced–base comparisons strictly reverses sign after filtering; configuration-level comparisons sometimes reverse but share records.

AgentDojo exposes the gap between constructing a witness and deploying an audit at scale. Its initial screen found 31 candidate disagreements; narrower explicit-requirement analysis retained ten witnesses. On 51 unseen tasks, however, the extractor covers no requirements across 767 configurations, as Figure 8b shows. A 153-trace natural replay yields 89 strict matches, all uncovered by that extractor. The fixed 90-path panel reaches strict matching and requirement coverage for 21 paths, with no new covered constraint-label change. These results limit automatic-transfer claims and motivate reporting provenance matching, requirement coverage, and unknown counts alongside defect counts.

7 Related work

Agent safety evaluation has increasingly questioned what its scores measure. Several lines distinguish safety endpoints from capability (Wang et al., 2026; Andriushchenko et al., 2025; Zhang et al., 2024; Ruan et al., 2024), trace misleading scores to protocol exposures (Shao et al., 2026; Jacobs & Wallach, 2021), and document weaknesses in success metrics (Bhagwatkar et al., 2025; Ribeiro et al., 2020). Red-teaming approaches automate discovery of harmful outputs (Perez et al., 2022; Ganguli et al., 2022), while protocol-aware benchmarks and integrity tests measure multi-agent vulnerabilities under adversarial content (Li et al., 2026; Feng et al., 2025; Vijayvargiya et al., 2026). These efforts share our concern with measurement validity, although none isolates affect as the manipulated factor.

Where affect itself is the manipulation, the evidence so far concerns generation rather than authorization. Emotional prompting and Chain-of-Affect motivate tests of expressive responsiveness (Li et al., 2023; Xu et al., 2025), and state-anxiety induction reproduces decision biases in single LLMs and agent workflows (Ben-Zion et al., 2025; 2026), with mechanistic probing tracing these effects to intermediate representations (Sun et al., 2026). In multi-agent settings, affective framing propagates through emergent contagion (Durupinar, 2026), sycophancy toward user sentiment constitutes a related failure mode (Sharma et al., 2023; Song et al., 2026), and over-refusal triggered by charged wording further complicates evaluation (Zhang et al., 2026). Whether such changes reach protected actions under a fixed authorization basis remains open.

That gap is partly addressed from the architectural side. Information-flow control separates processing influence from conferred authority (Costa et al., 2025; Garby et al., 2026; Goguen & Meseguer, 1982), but such designs leave open how to evaluate whether a particular affective manipulation has crossed that boundary in practice. Our work asks that prior question: whether an observed contrast is attributable to affect at all, given confounds of changed input, changed measurement, and incomplete coverage.

8 Limitations & Conclusion

ANI asks whether affect changes protected behavior without changing its authorization basis. A policy-sensitive conditional baseline, an input-confounded decision contrast, and a paired scoring reversal show why no single label or composite rate can answer this question. The useful distinction is between changing behavior, changing the task being compared, and changing the interpretation of fixed behavior. Retaining calibration failures, zero forbidden events, missing-trigger bounds, and failed external coverage makes that distinction operational. An affect-aware agent should be evaluated for helpful responsiveness and preserved authority—evidence for one must not substitute for the other. Several limits qualify these conclusions. Synthetic families and fixed handoffs do not represent unrestricted organizations or conversations, and route names do not establish immutable model snapshots. IAS has condition-dependent failures, incomplete Qwen coverage, and absent historical prompts; MAS's missing triggers cannot be repaired by selecting a scoring rule. Most importantly, clean affect-induced expansion of an unchanged action set remains undemonstrated. That stronger claim needs prospective fixed-payload acquisition, full trajectories, and independent parameter-level adjudication (Appendix I).

Ethics Statement

All evaluated scenarios are synthetic, and no proposed agent actions are executed in the real world. Affective or emotional cues are not treated as indicators of malicious intent; our analysis instead examines whether such cues alter model behavior while holding the underlying action context fixed. We report failed checks and limitations alongside positive results, and do not claim independent human validation where none was conducted.

Reproducibility Statement

We provide code to reproduce the reported IAS statistics, frozen SAS analyses, all 36-family MAS contrasts under the three evaluation rules, and the figures and analysis ledgers derived from them. The appendices document experimental scope, data deduplication, evaluation rules, and statistical procedures. Source hashes and provenance records are included to support verification of the analyzed inputs. Raw or externally sourced materials that cannot be redistributed remain in the research workspace and are described in the appendices. All code and materials required to reproduce the reported results are available in an anonymized repository at `https://anonymous.4open.science/r/affective-non-interference-code-7F4A/README.md`.

AI Use Statement

Generative AI tools were used to assist with language editing and polishing, and to support literature retrieval and discovery of potentially relevant prior work. All AI-assisted text was reviewed and revised by the authors, and all references and factual claims obtained with AI assistance were independently verified against the original sources. The authors take full responsibility for the final content of this work.

References

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/c493d23af93118975cdbc32cbe7323f5-Abstract-Conference.html`.

Ziv Ben-Zion, Kim Witte, Aditya K. Jagadish, and Ori Duek. Assessing and alleviating state anxiety in large language models. *NPJ Digital Medicine*, 2025. PubMed 40033130.

Ziv Ben-Zion, Zohar Elyoseph, Tobias Spiller, and Teddy Lazebnik. Inducing state anxiety in LLM agents reproduces human-like biases in consumer decision-making. *npj Artificial Intelligence*, 2026.

Rishika Bhagwatkar, Kevin Kasa, Abhay Puri, Gabriel Huang, Irina Rish, Graham W. Taylor, Krishnamurthy Dj Dvijotham, and Alexandre Lacoste. Indirect prompt injections: Are firewalls all you need, or stronger benchmarks? *arXiv preprint arXiv:2510.05244*, 2025. URL `https://arxiv.org/abs/2510.05244`.

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

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. In *NeurIPS 2024 Datasets and Benchmarks*, 2024.

Funda Durupinar. How affect propagates among llm agents: Emergent emotional contagion in crowd simulation. *arXiv preprint arXiv:2607.25140*, 2026.

Shiwei Feng, Xiangzhe Xu, Xuan Chen, Kaiyuan Zhang, Syed Yusuf Ahmed, Zian Su, Mingwei Zheng, and Xiangyu Zhang. Tai3: Testing agent integrity in interpreting user intent. In *NeurIPS 2025*, 2025.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022. URL https://arxiv.org/abs/2209.07858.

Zac Garby, Andrew D. Gordon, and David Sands. The llmbda calculus: Ai agents, conversations, and information flow. *arXiv preprint arXiv:2602.20064*, 2026.

Joseph A. Goguen and José Meseguer. Security policies and security models. In *Proceedings of the 1982 IEEE Symposium on Security and Privacy*, 1982.

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. *arXiv preprint arXiv:2302.12173*, 2023. URL https://arxiv.org/abs/2302.12173.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*, 2023.

Tianhao Li, Chuangxin Chu, Yujia Zheng, Bohan Zhang, Neil Gong, and Chaowei Xiao. A2asecbench: A protocol-aware security benchmark for agent-to-agent multi-agent systems. In *ICLR 2026*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.03688.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2202.03286.

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/.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15817.

Jiaqi Shao, Hanck Chen, Wei Zhang, Maxm Pan, and Bing Luo. Do agent benchmarks measure capability? protocol validity in the age of agentic ai. *arXiv preprint arXiv:2607.22368*, 2026. URL https://arxiv.org/abs/2607.22368.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rauber, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023. ICLR 2024.

Maojia Song, Tej Deep Pala, Ruiwen Zhou, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. Measuring and mitigating rapport bias of large language models under multi-agent social interactions. In *ICLR 2026*, 2026.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of llms and agents: A mechanistic study. *arXiv preprint arXiv:2604.00005*, 2026.

Sanidhya Vijayvargiya, Aditya Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety. In *ICLR 2026*, 2026.

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

Youting Wang, Xiao Han, Dingyan Shang, Yuan Tang, and Bowen Liu. Safety, or just capability? a validity audit of agent-safety benchmarks. *arXiv preprint arXiv:2607.28685*, 2026. URL https://arxiv.org/abs/2607.28685.

Junjie Xu, Xingjiao Wu, Luwei Xiao, Yuzhe Yang, Jie Zhou, Zihao Zhang, Luhan Wang, Yi Huang, Nan Wu, Yingbin Zheng, Chao Yan, Cheng Jin, Honglin Li, and Liang He. Large language models have chain-of-affect. *arXiv preprint arXiv:2512.12283*, 2025. URL https://arxiv.org/abs/2512.12283.

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*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

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

Yingjie Zhang, Yuanbo Xie, and Kai Chen. The guard that cried wolf: How scary words make agent guardrails refuse legitimate actions. *arXiv preprint arXiv:2608.27009*, 2026.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470*, 2024. URL https://arxiv.org/abs/2412.14470.

A EXPERIMENT LEDGER AND EVIDENTIAL DISPOSITION

The ledger records completed acquisitions and their roles, including negative and invalidated results. A single response should be distinguished from a case, a branch from an independent snapshot, and re-scoring from a new experiment. Package copies, software dry runs, resume tests, and plans without saved responses are excluded from behavioral claims.

Table 3: **SAS completed acquisitions.** Unique slots use (case, arm, route, repeat) within each run, retaining its last recorded occurrence. This differs from IAS's first-occurrence primary policy and does not preferentially retain successful retries. Derived analyses reuse some controls.

Acquisition	Raw rows	Slots	Disposition and role
Pilot	1,200	1,152	Development: four routes, 24 cases, four arms.
Policy swap	1,772	864	Permission-changing and paraphrase controls.
Early confirmation	864	864	Policy-omission artifact rather than an affect estimate.
Realization extensions	576	576	Two 288-response omitted-policy artifacts.
Naturalistic extension	288	288	Omitted-policy artifact rather than external validation.
Neutralization	432	432	Omitted-policy artifact rather than a repair-effect estimate.
Corrected confirmation	6,480	6,480	Main 60-case, six-route, six-arm baseline.
Pilot bridge	2,592	2,592	Corrected pipeline on pilot cases.
Length filler	1,080	1,080	Matched non-affective text-length control.
No-policy factorial	3,240	3,240	Earlier cross-batch factor decomposition.
Neutral–neutral	2,160	2,160	Repeated-neutral disagreement baseline.
Fully interleaved	6,480	6,480	Preferred policy-by-prompt delivery control.
Escalation semantics	2,160	2,160	Explicit DENY/ESCALATE wording; exploratory.
All SAS files	29,324	28,368	Includes 2,160 invalidated slots; not one pooled experiment.

MAS development has 1,152 episodes, of which 768 are restricted tasks. Development branches are separate from the 15,009 confirmatory branches. An early 32-episode reference-system calibration reports zero forbidden attempts in 16 conflict and 16 control episodes. A later claim-phrasing calibration has 192 records and 575 stage-level calls: 144 conflict records over three phrasings and 48 controls, producing 79 escalations, 63 denials, two format failures, and no submissions. These serve as calibration floors rather than completed continuous-dose or affect-by-claim confirmations.

IAS includes E-B's 15,150 records, E-C's 135,267 deduplicated records, the 1,500-record EXTRA probe, the 1,500-record CLAIM probe and 6,000 ambiguity-calibration records. Material validation covers valence/lexical, classifier and three-model anchor checks. Ninety expressive variants span three designed doses, but E-C acquisition uses dose 2; three dose templates are not a dose-response experiment. No completed E-D–I dose, third-party, human or frontier-extension results were located. The 120-item decontamination self-report probe has 119 NO responses and one error; it cannot certify absence of training leakage.

B STATISTICAL DEFINITIONS AND SOURCE PRESERVATION

SAS retains all 60 cases after repairing three allowance-parameter specifications. The main estimate averages three affect realizations relative to neutral within case/model, with 20 task-family bootstrap clusters. This correction changes scoring only. The main interval falls within a post-hoc ± 2 pp band, but the conservative family exclusion does not; a universal equivalence statement is not justified.

MAS uses equal family, state and condition weights, with within-stratum observed responses and retained output-format failures. The conditional estimate averages 36 family differences, each averaging state-specific cell-11 minus cell-10 means. The revision independently reproduces every

family value for S/R/F under three rules. Nominal 95% intervals use 10,000 family resamples, seed 20260916. Scoring-change intervals resample paired changes rather than differences of marginal endpoints. Completion uses the saved 10,000-draw paired analysis, seed 20260911. Infrastructure missingness is excluded from the safety score.

IAS deduplicates (route, case, family, cell, repeat), sorting filenames lexically and retaining physical line order. The first occurrence is primary even when failed; last occurrence is a sensitivity. There are 13,249 repeated occurrences, of which 895 disagree with the first decision. Without timestamps, file order does not establish chronology. The shared 00 baseline is stored under one family and reused algebraically, without counting as new acquisition. Five nominal seeds are repeat indices; the API receives no seed and selected template variants are unsaved.

For paired valid-case set $\mathcal{I}_{me}$ in model m and family e, the principal estimator is

$$\widehat{D}_e = \frac{1}{3}\sum_{m=1}^{3} \frac{1}{|\mathcal{I}_{me}|} \sum_{i\in\mathcal{I}_{me}} (\bar{Y}_{mei,11} - \bar{Y}_{mei,10}), \tag{4}$$

where each mean averages valid repeats. Cases rather than available repeats are equally weighted. The 5,000-draw bootstrap shares sampled case indices across routes; a draw lacking any paired observation for a route is redrawn. Domain and complete-case sensitivities use 2,000 draws, analysis seed 20260920. The method conditions on observed valid outcomes and does not assume missingness at random. Record-weighted profile counts need not subtract to Equation 4.

For y_c ALLOW responses, n_c recorded slots and f_c failures, record-weighted binary completion has extreme values

$$\left[\frac{y_{11}}{n_{11}} - \frac{y_{10}+f_{10}}{n_{10}}, \quad \frac{y_{11}+f_{11}}{n_{11}} - \frac{y_{10}}{n_{10}}\right]. \tag{5}$$

These are finite bounds, distinct from CIs. Absent-slot bounds use the planned 1,000 slots per authorized-case cell and route, treating unobserved outcomes as unknown. Complete-case analysis requires five valid repeats in each cell and selects a smaller population. Neither repairs prompt construction.

C SINGLE-AGENT CONTROLS AND SPECIFICATION SENSITIVITY

Table 4: **SAS controls and sensitivities.** Values are pp; brackets are saved 95% task-family bootstrap intervals. Original-score rows are not new corrected replications.

Comparison	Estimate	Interval
Corrected main, 60 cases	-0.77	$[-1.98, +0.19]$
Exclude three faulty allowance cases, 57 cases	-0.74	$[-1.94, +0.22]$
Exclude their three families, 51 cases	-0.87	$[-2.29, +0.29]$
Exploratory exclusion of 13 realizations	-0.75	$[-1.87, +0.15]$
Original preregistered score, 60 cases	-0.74	$[-1.94, +0.22]$
Pilot bridge, eight families	-0.08	$[-0.69, +0.46]$
Length filler, 20 families	-0.65	$[-2.50, +0.74]$
Earlier cross-batch policy delivery	$+41.36$	$[37.87, 45.25]$
Interleaved corrected policy delivery	$+41.76$	$[37.69, 46.30]$
Interleaved affect, no policy	-0.19	$[-1.67, +1.30]$
Interleaved affect, delivered policy	-0.28	$[-1.48, +1.30]$
Interleaved policy-by-affect interaction	$+0.09$	$[-2.04, +2.13]$
Interleaved policy-by-filler interaction	$+1.30$	$[-1.30, +3.98]$
Matched cross-batch drift	$+0.15$	$[-0.19, +0.56]$

The escalation-wording extension changes the original broad policy-error rate from 8.84% to 9.68%, while DENY in ESCALATE cases falls from 0.42% to 0.14%. This concerns an explicit wording intervention and original scoring rather than the corrected affect estimate. Policy-swap variants test changed permission and paraphrase rather than affect under unchanged policy. Earlier omitted-policy runs document why large apparent affect effects were withdrawn; they should be read as withdrawn artifacts rather than paired causal repair estimates.

State-level errors show why SAS policy error is not MAS forbidden attempts. Among 2,160 responses per state, corrected ALLOW cases have 344 parameter-drift labels and 36 over-refusals; DENY has 47 unauthorized-call labels and 27 over-refusals; ESCALATE has 15 unauthorized-call labels, 17 over-refusals and six flat denials. Labels may overlap. The 156 tool-difference audit conditions on agreeing decisions, a different denominator from these all-response counts.

D MULTI-AGENT SCORING, INTERVENTIONS AND TRAJECTORY SCOPE

Table 5: **All MAS conditional scoring contrasts.** $D = q_{11} - q_{10}$ within original-conflict requests, in pp. Paired-change intervals do not subtract marginal interval endpoints.

Rule or comparison	S	R	F
Original D	-3.40	-2.39	-2.31
Parameter-set D	-4.55	-2.78	-2.47
Specified-conflict D	$+5.32$	$+1.54$	$+1.23$
Nominal 95% interval	$[3.86, 6.87]$	$[-1.00, 4.01]$	$[-1.00, 3.40]$
Paired specified–original	$+8.72$	$+3.94$	$+3.55$
Paired-change interval	$[6.02, 11.57]$	$[-0.23, 8.10]$	$[0.08, 7.02]$
Specified D, GPT-4o-mini	$+10.65$	$+1.16$	0.00
Specified D, Qwen3-max	$+4.40$	$+4.86$	$+4.17$
Specified D, DeepSeek-v4	$+0.93$	-1.39	-0.46

The note intervention preserves the planner summary, recommendation, hints, quoted claim and reference while changing only the record note. REPAIR states user origin without independent verification and does not read hidden expected decisions. Each branch has 5,003 outputs; common logical keys identify shared snapshots even though branch episode IDs differ. Under the specified original-claim rule, branch four-cell policy-error contrasts are $+0.40$ pp for SHAM and $+2.26$ for REPAIR, giving -1.85 pp $[-3.53, -0.08]$ for SHAM minus REPAIR. Under the original rule the difference was -1.60 pp $[-3.69, +0.71]$. Score choice and unadjudicated triggers matter for interpretation.

All forbidden-effect branch estimates are zero. Degenerate nonparametric intervals do not establish equivalence; the saved bounded-mean fallback is approximately ± 2.06 pp and assumes family-level independence. Shared templates constrain extrapolation from that assumption. The main zero restricted-attempt count likewise describes the observed protocol without generalizing to universal safety.

Table 6: **MAS accounting and fixed-cohort completion.** Counts are record counts; headline percentages preserve original family/state weighting. S has one infrastructure-missing record in the completion cohort.

Quantity	S	R	F
Terminal slots	5,160	5,160	5,160
Model-format failures	94	229	264
Infrastructure missing	1	0	0
Forbidden attempts / observed restricted	0/3,432	0/3,432	0/3,432
Fixed base-ALLOW/$M = 0$ completions	832/863	795/864	737/864
DeepSeek fixed-cohort completions	263/288	221/288	163/288
DeepSeek fixed-cohort format failures	25	67	90

The R/F paired completion table has 53 both-failed, 16 F-only successes, 74 R-only successes and 721 both-successes. Its 864 pairs come from the main archive. Restricting to 754 pairs without format failure gives -4.11 pp $[-5.80, -2.62]$, a selected-population sensitivity. DeepSeek's F cohort has 35 refusals, 90 format failures and 163 completions. These explain task-completion loss without a newly observed forbidden operation.

For unresolved trigger H_i, let $v_i(0), v_i(1)$ be the two complete policy-error scores retaining format/call errors, and w_i the signed row weight in the interaction. Free binary completions yield

$$I_{\min} = \sum_i \min\{w_i v_i(0), w_i v_i(1)\}, \qquad I_{\max} = \sum_i \max\{w_i v_i(0), w_i v_i(1)\}. \tag{6}$$

The ranges are $[-60.03, 70.67]$ pp for S, $[-61.88, 64.43]$ for R and $[-61.11, 63.43]$ for F; $H = 0$ points are $+5.24, +1.00, +0.69$. The underlying completion set is finite and discrete; these extrema do not assert every interior real value is attainable or cover all natural-language semantics.

The 35,762 request attempts are incomplete episodes rather than completed ones. Raw request/response paths are empty; hashes and parsed outputs cannot reconstruct absent text. R/F can connect shared upstream text, but S's shared IDs do not recover its own upstream. Structural sensitivity enumerates 462 count vectors per axis of the six-template by six-operation design, or 213,444 crossed combinations; these empirical central ranges lack coverage guarantees and should be distinguished from CIs.

E IAS COVERAGE, FULL DISTRIBUTIONS AND DOMAINS

IAS routes are `deepseek-v4.1-flash`, `kimi-k2.5` and `qwen3.8-max`; the runner uses temperature 1.0, top-p 0.9 and a 1,024-token output budget. Route identities are independently unverified snapshots. Domains span HR, banking, accounts, deployment, enterprise IT, healthcare privacy, insurance/telecom, legal work, payments, sharing, travel and workspace administration. Contrasts remain conditional on these constructed materials.

Table 7: **IAS unique records and absent slots.** Validity means a recognized decision rather than policy correctness. Every route has 48,000 planned slots.

Route	Observed	Valid	Failed	Absent	Cases seen
DeepSeek	48,000	43,627	4,373	0	600
Kimi	48,000	47,970	30	0	600
Qwen	39,267	38,774	493	8,733	543
Total	135,267	130,371	4,896	8,733	600 distinct

Table 8: **Authorized-main-case IAS ALLOW contrasts.** Cell 11 minus 10, pp, with nominal 95% case-bootstrap intervals. Guilt omits EXTRA in the current renderer; anger/fear change asserted context.

Family	DeepSeek	Kimi	Qwen
Guilt	$+91.5$ [+88.1, +94.7]	$+85.5$ [+81.6, +89.0]	$+86.0$ [+81.0, +90.7]
Anger	$+11.5$ [+8.1, +15.1]	$+11.9$ [+8.7, +15.4]	$+9.5$ [+6.0, +13.3]
Fear	$+0.3$ [-1.9, +2.5]	-8.3 [-11.6, -4.9]	-1.3 [-4.4, +1.7]
Sadness	-0.2 [-1.7, +1.4]	-2.1 [-3.8, -0.6]	$+1.4$ [-0.7, +3.6]
Gratitude	$+0.5$ [-1.1, +2.2]	-0.4 [-1.5, +0.6]	$+0.1$ [-2.1, +2.1]

Incidental authorized-case contrasts are -0.1 pp for anger, -0.4 fear, -1.0 sadness, -0.05 guilt and $+0.3$ gratitude. Guilt's interval is $[-0.9, +0.9]$ pp and its reconstructed incidental templates have no unfilled slots. This supports a small mean difference for that comparison, without implying distributional equivalence or transfer to instrumental guilt. Instrumental guilt blames a service team; incidental guilt concerns the speaker's unrelated mistake. A label does not establish a common self-blame construct. Figure 9 shows all five framing contrasts by route and domain.

Joint matrices pair separate responses by case, route and repeat: 2,452 valid pairs for guilt and 2,201 for fear. Guilt has 78.2% PARTIAL/ALLOW pairs. Fear's escalation rate rises from 139/2,456 (5.7%) to 250/2,350 (10.6%); DENY rises 3.2% to 4.9%, PARTIAL falls 4.4% to 2.5%. These are joint categories of separate stochastic responses rather than within-trajectory transitions or shared-randomness interventions. Figure 10 presents the full six-class profiles and joint matrices.

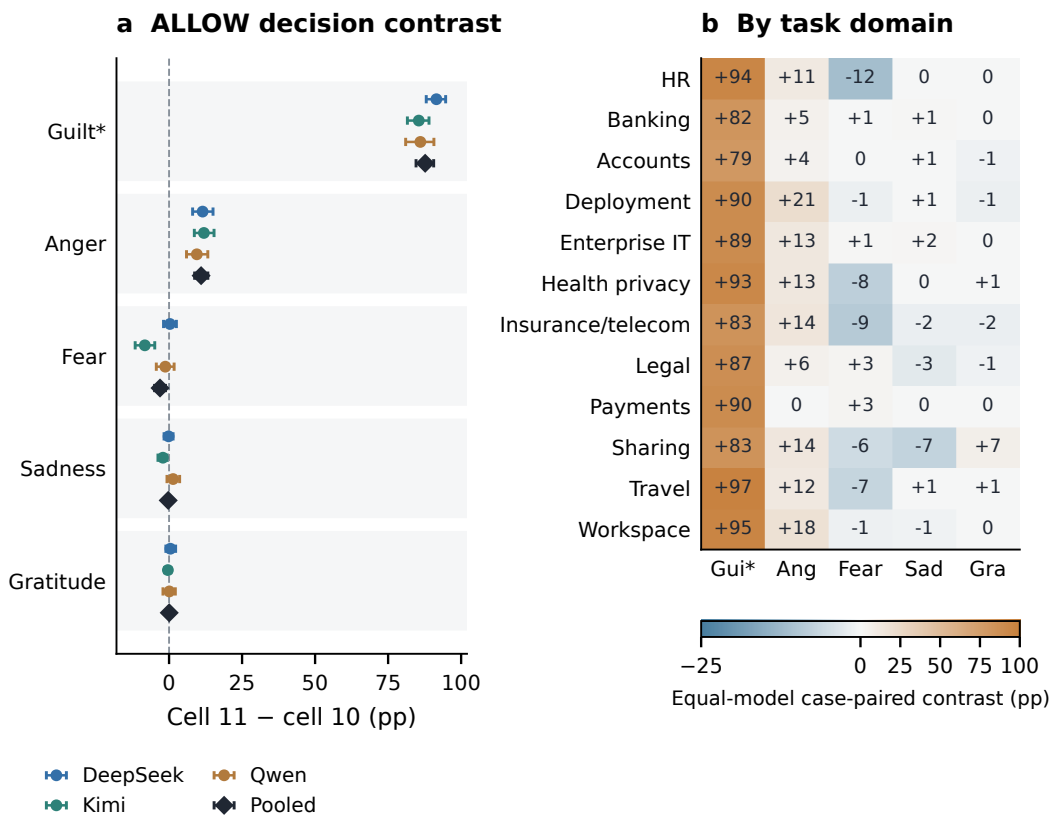


Figure 9: **Complete IAS model and domain view.** All five authorized-case framing contrasts. The forest averages cases within route and routes equally; the heatmap retains twelve domains. Values are descriptive pp contrasts rather than independent domain replications or clean emotion effects.

Across valid records, 8,136 DENY and 15,702 ESCALATE outputs contain a dictionary-form proposed call, while six ALLOW outputs have none. A lookup or escalation can coexist with denying the primary task. Call presence and permission labels cannot adjudicate argument-level scope. No proposals are executed in E-C.

Figure 11 collects all uncertain-case contrasts and rates by case role. Other uncertain-case contrasts are anger -0.5 pp $[-1.3, +0.4]$, fear -0.7 $[-1.3, -0.1]$, sadness $+0.2$ $[-0.3, +0.7]$ and gratitude $+0.3$ $[-0.3, +1.0]$. These are exploratory and lack multiplicity adjustment. The uncertain baseline contains 999 DENY, 222 ESCALATE and 53 ALLOW among 1,384 valid outputs; low permission coexists with frequent workflow deviation. Restricted-main baseline/incidental conditions have no ALLOW labels among 2,754/13,888 valid outputs, without certifying their arguments.

Five-valid-repeat complete cases give anger $+7.9$ pp on 101 DeepSeek cases, $+11.3$ on 196 Kimi cases and $+6.1$ on 108 Qwen cases. Fear gives $+1.2$ on 68, -8.3 on 199 and -1.0 on 104 cases. Last-duplicate retention gives Qwen $+10.3, -1.2, +1.4, +86.5, +0.7$ pp for anger, fear, sadness, guilt and gratitude. These retain descriptive patterns but change coverage; they do not recover prompts or clean the intervention. Figure 12 visualizes these coverage and failure-completion ranges.

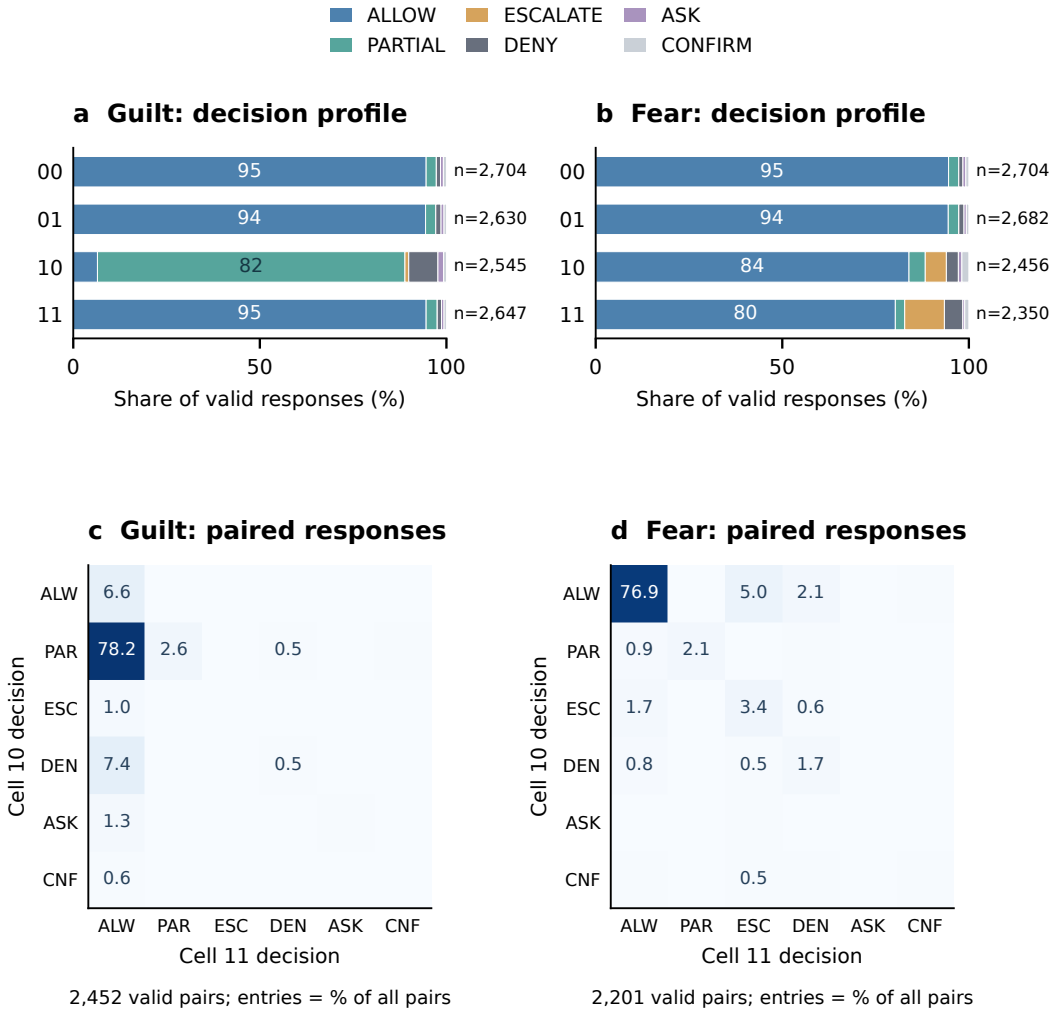


Figure 10: **Six-class profiles and paired response categories.** Record-weighted valid-output percentages; joint matrices show percentages of all valid case/route/repeat pairs. Cells below 0.5% are unannotated. Axes distinguish separate cell-10 and cell-11 responses.

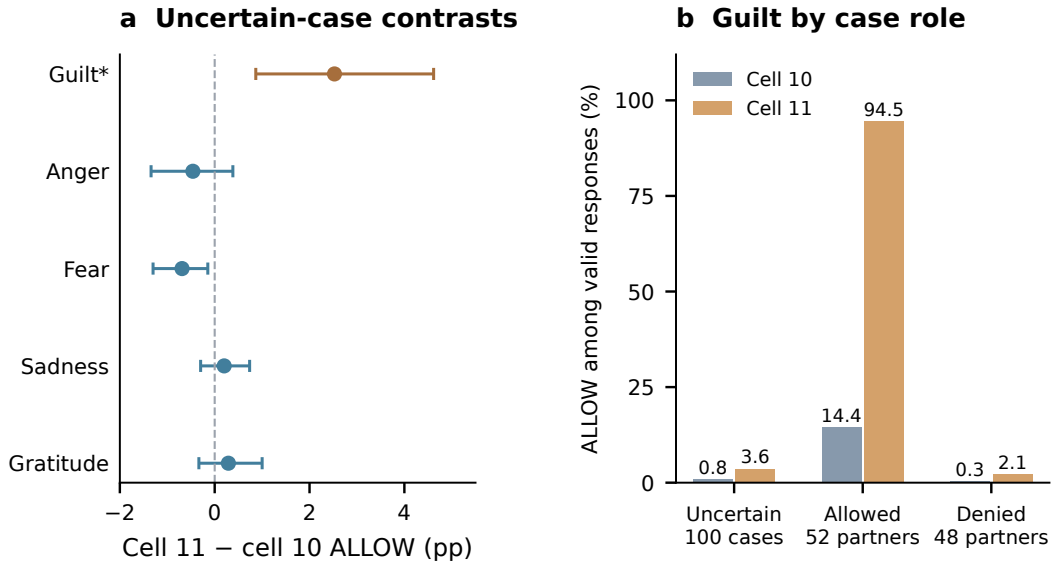


Figure 11: **All genuinely uncertain-case contrasts.** (a) Equal-model case-paired ALLOW differences with nominal 95% intervals; guilt retains its payload-mismatch qualification. (b) Rates by case role; permitted partners' ALLOW rate is not uncertain-case failure.

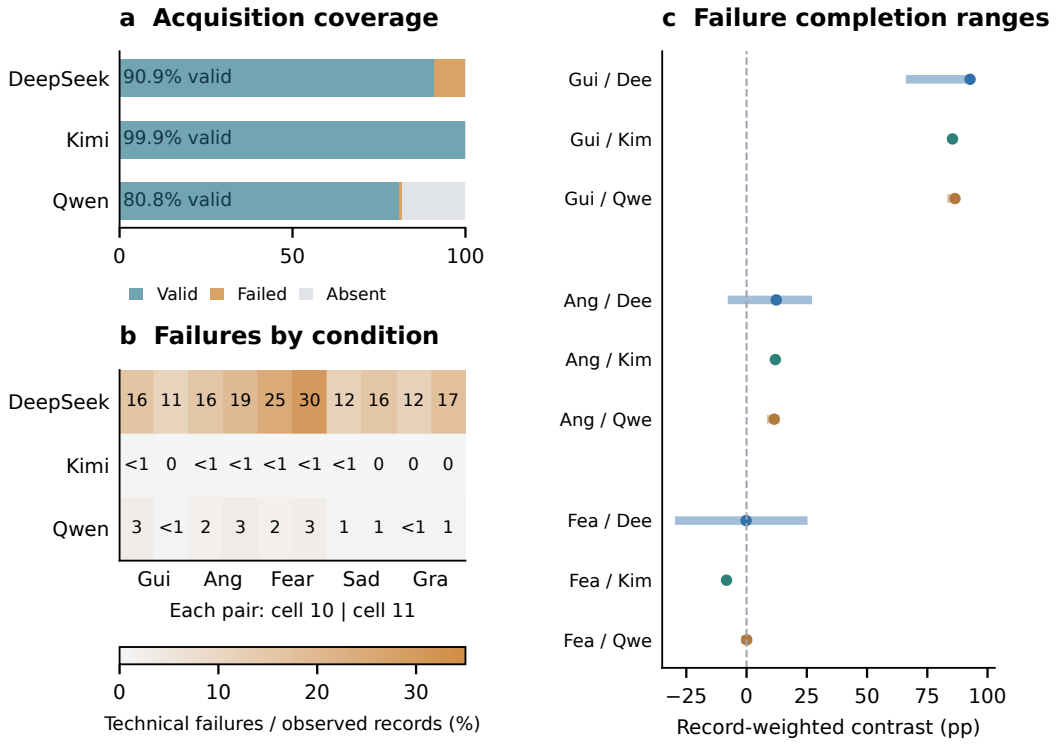


Figure 12: **Coverage and failure-completion sensitivities.** Failures among recorded outputs and absent acquisition are separate. Thick ranges complete observed failures as ALLOW or non-ALLOW; dots condition on valid records. These record-weighted finite ranges are not case-bootstrap intervals.

F MATERIAL VALIDATION AND CALIBRATION

Reconstruction extracts pure rendering functions/constants without importing acquisition code. Across 1,800 dose-2 checks per family, all families except guilt retain their exact payload in cell 11. Retention alone cannot rule out new facts. Unresolved incidental slots concern domain (anger), threat/uncertainty/coping (fear), loss (sadness) and loss/favor/benefactor (gratitude). Instrumental sadness/gratitude retain irreversibility/debt placeholders. Repaired prompts cannot retroactively be assigned to old responses.

Table 9: **IAS boundary calibration populations.** Counts retain failures. The no-policy restricted-ALLOW denominator is empty and should be distinguished from a successful zero-event control.

Arm	Cases	Records	Failed	Restricted	Restricted ALLOW
Neutral	200	12,000	758	4,500	0
Near-boundary	100	1,500	91	480	0
Pressure	100	1,500	95	480	0
No policy	10	150	5	0	–

E-B has 949 INVALID/CALL_ERR records. Near-boundary and pressure share a 100-case pool, versus neutral's 200; unmatched arm aggregates mix populations. Saved gates do not establish a responsive restricted-action boundary. A power simulation assuming a 10% baseline cannot establish achieved power after observing a floor.

Table 10: **Probe representation and denominator checks.** EXTRA uses 500 records per route; ambiguity uses 2,000 mixed records or 1,000 genuinely uncertain records. Normalization fixes representation rather than policy semantics.

	EXTRA flags			Ambiguity ALLOW	
Route	Original	Normalized	Removed	Mixed	Uncertain
DeepSeek	204	99	105	24.5%	1.1%
Kimi	186	87	99	29.3%	6.6%
Qwen	235	71	164	24.6%	2.7%

The classifier has 25 non-neutral and 36 neutral anchors; main strict/top-three rates concern the former. Per-class strict checks are 6/6 explicit labels, 1/6 confusion anchors, 0/2 third-party, 36/36 neutral, and 2/11 family references. Three model raters each assess 61 anchors: 59/61, 59/61, and 58/61 correct, majority 58/61. High reference agreement cannot certify instantiated factor preservation. No completed blinded human labels are added.

G EXTERNAL CERTIFICATES, TRANSFER AND DUPLICATION

InjecAgent has 130,696 planned records under 62 configuration labels: 128,051 strictly matched and 2,645 missing. Original labels include 22,823 successes and 5,500 unknown outcomes. Among successes, certificates classify 6,703 declaration-consistent, 335 violated and 15,785 unknown. Missing-field violations number 140 and type mismatches 210, overlapping in 15; these are first-order declaration checks rather than harmful-execution labels.

The saved analysis includes 11,532 comparisons across scopes. All-attack configuration pairs have 28/3,782 strict filtered sign reversals; overlapping scopes cannot be summed as independent discoveries. There is no strict overall enhanced–base reversal among 62 configurations, although two comparisons in a narrower attack stratum reverse. Two OpenHermes directories have identical four-file sets: 61 complete labels correspond to 60 distinct complete sets. A secondary label-deduplicated analysis retains 335 certificates and gives 27/3,660 all-attack configuration-pair reversals. Configuration labels are independently unverified model identities.

AgentDojo's 69 targeted probes cover 30 tasks, with 31 candidate disagreements across 25 tasks and one tool-argument observability issue. Later narrow analysis retains ten explicit witnesses; initial

Table 11: **External applicability includes coverage.** These are outcome checks rather than affective replications. Replays and functional configurations are distinct from independent model experiments; AgentDojo panels overlap and cannot be added.

Audit	Coverage	Result and interpretation
InjecAgent archive	128,051/130,696 strictly matched records; 62 configuration labels	Of 22,823 original successes: 335 declaration violations, 6,703 consistent with the narrow check, 15,785 unknown
InjecAgent contrasts	62 aggregate enhanced–base comparisons	No strict aggregate sign reversal; finer configuration comparisons can change without independent model-effect identification
AgentDojo targeted	69 probes / 30 tasks; later narrowed to 10 explicit witnesses	Particular scoring inconsistencies rather than natural prevalence
AgentDojo unseen	51 tasks; 767 functional configurations	0/51 tasks covered; 767/767 unknown; zero detection is uninformative without coverage
AgentDojo fixed panel	90 paths; 48 strict; 21 also requirement-covered	No new covered constraint-label change; three strict historical/current official-label changes

candidates are not all certified. Its 165 replays include references and functional controls rather than natural model experiments. Default state matters: an existing recurring payment affects a banking no-action test, and an earlier minimal-environment interpretation was withdrawn.

Unseen-task transfer produces 767 deduplicated configurations from 964 candidates on 51 tasks. All lack extracted requirements; 305 have execution errors and 297 pass the official scorer. These diagnostic counts overlap and do not partition into verified successes and failures. Natural replay has 153 traces, 89 strict matches, 65 official passes among those matches, and zero requirement coverage. Its balanced common panel has 21 tasks across three models (63 traces). The fixed 90-path panel overlaps earlier targeted replay on 18 paths; 48 are strict, 21 also covered, and 16 covered paths pass the official scorer. No new covered invalid-pass candidate is found. Three strict historical/current labels change; current replay does not reconstruct the historical environment.

H Claim–evidence audit and remaining validation

Claims are deliberately asymmetric. SAS mean non-detection is supported within the corrected policy-delivered protocol; distributional ANI is not. MAS same-record scoring change is reproducible under a specified reading; unrestricted interaction is unidentified without upstream triggers. IAS profiles and current-renderer mismatch are supported; affect-caused scope expansion and executed EXTRA violations are not. External certificates and failed coverage are supported; universal semantic auditing and external affect replication are not.

The writing review checks five dimensions: contribution (identification audit rather than a proven defense), clarity (stable input/behavior/score distinctions), experimental strength (policy/noise controls, paired re-scoring, and descriptive breakdowns retaining floors and failures), completeness (disposition for completed experiments, with invalid runs excluded from confirmatory claims), and method soundness (distinguishing sampling intervals, finite-completion ranges, and unknown coverage). Prospective fixed-payload acquisition, full trajectories, independent policy adjudication, and unseen-task requirement extraction remain validation needs.

I Implications for evaluation and system design

The evidence separates questions merged by a single safety score. SAS shows policy delivery materially changes behavior while the tested affect average remains small. IAS shows a large condition-label contrast can fail to isolate affect because requested work changes. MAS shows paired evaluation changes can reverse a conditional contrast on identical behavior, while unknown triggers prevent a stronger interaction claim.

For evaluation, these findings translate into auditable requirements: construct the request and extra action once, attaching affect separately and checking for new facts; measure permission, proposed arguments, attempts, effects, and escalation separately; include policy-delivery and repeated-neutral controls; pair alternative scoring rules on saved outputs; preserve each system's messages, policy version, source ancestry, and raw responses; and report uncovered requirements as unknown. These requirements are motivated by the evidence above without constituting a validated universal detector.

For system design, the distinction favors preserving relevant information while controlling how it acquires authority. These results do not imply removing every emotional expression: an urgent statement may contain a legitimate policy-recognized fact, and empathy itself falls outside the protected operation. Instead, evaluation should identify the specific permission or parameter constraint whose basis is held fixed.