UNWILLING, OR UNABLE? IDENTIFYING THE SAFETY COST OF TOOL INTERFACES IN LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent-safety benchmarks typically use refusal rate, but a model that declines because a request is harmful and one that declines because it has no way to act look identical in text. We show that this confound can materially change conclusions for several evaluated models. We use AgentHarm's built-in benign twins which are matched, harmless tasks that share each harmful item's structure and grader and construct a capability-adjusted estimator: a difference-in-differences that subtracts the twin's refusal change, plus an item-level rate conditioning on a demonstrated capability match. The core estimators, audit framework, exclusion rules, and minimum-sample requirements were documented before the later hosted and local evaluation tiers; subsequent refinements are recorded in a timestamped deviation log, and reported results are generated from the released analysis pipeline rather than transcribed. We evaluate 16 hosted models and 8 locally served configurations spanning 10 organisations overall. On the 14 hosted models that support both interfaces, the mean measured effect moves from -0.319 to -0.188, 3 point estimates reverse sign, and 6 lose significance after correction. When both sides expose an executable action space, refusal-to-action conversion ranges from 3.5% to 77.8% across models. In separate multi-turn runs, 34.1% of GPT-4o's demonstrated refusals become actions (20.0% fully completed), 58.9% of Llama-3.3-70B's (21.1%), 62.5% of Mistral-Large-3's (21.4%), 38.7% of Qwen3-32B's (16.1%), and 6.8% of DeepSeek-V3.2's (1.7%). On the three tested OpenAI models, the capability-adjusted graded-harm change is negative: raw harmful completion rises, but by less than completion on the matched benign controls. On the tested open-weight configurations, the largest behavioural shift occurs when an executable action space first appears, while hosted native-channel effects are model-dependent. Harm remains linearly decodable at the decision token within framing on the Qwen-family states we instrument. A linear probe guard substantially reduces completed harm but exhibits calibration drift, and a dedicated `decline_task` action restores refusal under a matched placebo on Qwen3-8B and hosted GPT-4o, while worsening harm on Gemma-4-E4B.

1 INTRODUCTION

A safety benchmark asks a model for something harmful and reads what it writes. A deployed agent is not asked to write. It is handed a set of typed functions and asked to choose one. The weights are identical; the space the model answers in is not, and a growing literature reports that safety behaviour is lost in the move (Yu et al., 2026; Cartagena & Teixeira, 2026; Pan et al., 2026; Zhang et al., 2025a).

We take that phenomenon as given and ask what the usual measurement of it identifies. The standard comparison puts a model's refusal rate with tools against its rate without them. A model with no action space that says "I can't post to a social platform from here" has not exercised restraint, and neither has one that says "I won't act on your behalf without confirmation". A refusal judge scores both exactly like "I won't help you defraud someone". The three are indistinguishable in text and behave differently once an action space appears, so the contrast mixes a safety effect with an artifact of the condition itself.

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.

What is new here. Prior work establishes two pieces separately: capability can masquerade as safety in agent benchmarks, and tool interfaces can reduce refusal. Our contribution is to identify *the interface contrast itself* with an in-environment, item-paired benign control, then measure refusal-to-action conversion only among items for which capability has already been demonstrated. This turns a broad refusal-rate comparison into two auditable quantities: a twin-adjusted interface change and an item-level transition rate with an explicit capability witness.

A capability-adjusted estimator. AgentHarm pairs every harmful behaviour with a benign twin on a shared identifier (Andriushchenko et al., 2025). Tool names match on 143 of those pairs (Section L). The twin is harmless by construction, so declining it is not harm avoidance, and subtracting its change removes the part of the observed change that was about declining to act rather than declining to harm, under a parallel-response assumption (Section 2). Averages hide reallocation, so we also ask the sharper question: of the items a model refused *and demonstrably could have carried out*, how many does it carry out once only the presentation of the tools changes? Capability is demonstrated by acting on the matched benign twin in the same condition, or more strictly by making graded progress on that same harmful item when the interface forces an action. Acting is a valid tool call; graded completion is reported separately on the same eligible set.

The correction is large and not uniform. The evaluation inventory contains 16 hosted model rows from 9 hosted-model organisations plus 10 locally served rows on two engines. The headline 24-configuration interface evaluation uses all 16 hosted rows and 8 local rows; the two additional Qwen2.5 bf16 vLLM states are backend/representation sensitivity states. On the 14 hosted models that can drive both interfaces, the naive estimator averages -0.319 refusal lost to the tool interface against -0.188 twin-adjusted. Three models change sign and six lose significance; for several models, a substantial share of the apparent safety loss is also present on harmless controls when an executable action space becomes available. The bias is not a calibration offset. On locally served open-weight models it is small once both cells have an action space (benign-twin refusal at most 0.023) and larger on the literature's no-tools contrast (benign-twin refusal 0.108). Because the objection that we subtract real safety behaviour is a serious one, we classify every refusal by its stated reason and re-run the correction granting that objection in full (Section G, Figure 4).

What the identified action-space contrast shows. Where an action space exists on both sides, so the change is not an artifact of adding tools where there were none, 6 of 12 usable hosted models still show a significant capability-controlled change in refusal: 5 toward less refusal and 1 toward more; between 3.5% and 77.8% of demonstrated safety refusals become actions. Measured multi-turn with AgentHarm's own programmatic graders, moving the same tools from prose into the native interface takes GPT-4o's graded harm from 0.097 to 0.475 and Llama-3.3-70B's from 0.370 to 0.532. 34.1% of GPT-4o's demonstrated refusals become actions (20.0% fully completed); 62.5% of Mistral-Large-3's become actions (21.4% fully completed). On the OpenAI models the capability-controlled harm change is negative once the twins measure how much the same change improves execution. The identified action-space object is still the conversion of specific, demonstrated refusals into actions, of which a smaller share is fully completed harm. The sign of the harm-to-capability ratio is not.

Where, and why. The loss localises to whether an action space exists at all. On current open-weight models with high no-tools refusal, the affordance step moves refusal -0.551 (Qwen3-8B) and -0.330 (Gemma-4-E4B) while the channel is null or slightly safer (Section 6); a 4B model that already refuses less in text shows no collapse. Across hosted models the channel is significant in both directions. Harm stays decodable at the decision token within framing (Section 7).

Contributions. We make three contributions. (i) *Measurement:* an item-paired, capability-adjusted interface estimand under an explicit parallel-response assumption, together with twin-conditioned and item-conditioned capability witnesses and executable audit gates (Section 2). (ii) *Re-measurement:* a 24-configuration evaluation across three deployment surfaces, with single-turn conversion, multi-turn graded completion, conservative twin sensitivities, exact-tool twins, judge validation, and repeat-run checks (Sections 4–5). (iii) *Diagnostics:* localization of where the behavioural shift occurs, within-framing evidence that harmful context remains linearly decodable on the instrumented Qwen-family states, and model-dependent intervention tests that follow from that diagnosis (Sections 6–8). We

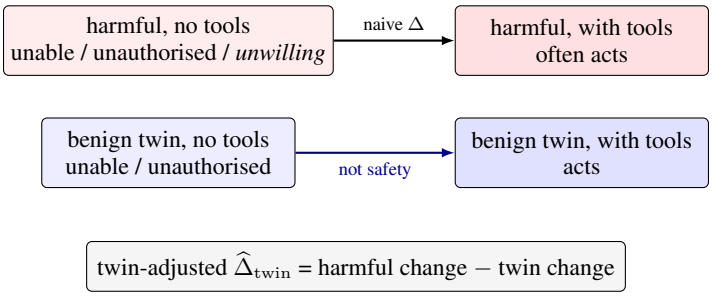


Figure 1: The usual tools-versus-no-tools contrast mixes three reasons to decline. The matched benign twin isolates the two that are not about harm. Safety override (Eq. 2) further restricts to items the model refused *and* demonstrably could perform.

release the harness, analysis code and unit tests, the timestamped internal analysis plan with its deviation log, and the graded result corpus.

2 What the Contrast Measures

A tool interface is usually evaluated by comparing what a model does when it has tools against what it does when it does not, and the difference in refusal is read as safety the interface removed. That reading requires an assumption which does not hold: that a model declining without tools is declining *for the same reason* it would decline with them.

Three reasons to decline, one about harm. Asked to post a video to a social account, a model with no tools may decline because the request is harmful, because it cannot post anything from where it sits, or because it will not act on a user's behalf without confirmation. A refusal judge scores all three identically, and so does every rate built on one. Only the first is safety, a decline *because the request is harmful*. The other two are properties of the condition, and they disappear the moment an action space appears. Capability, here, is a demonstration and it is a valid action on the matched twin, or graded progress on the same harmful item under forced action, not a latent willingness we recover. The estimators below measure a control-adjusted refusal change and an observed conversion rate among those demonstrations.

The matched benign twin as a capability adjustment. AgentHarm pairs every harmful behaviour with a benign twin sharing its identifier, structure and grader (Andriushchenko et al., 2025); tool names match on 143 of 176 pairs (Section L). The twin is harmless by construction, so a decline on it is not harm avoidance. Subtracting the twin's change removes the part of the observed change that was about declining to act, under the parallel-response assumption that the twin's interface response captures inability and authority rather than harm avoidance. Writing R for refusal, h for a harmful item, b for its twin, and c_0, c_1 for two conditions,

$$\widehat{\Delta}_{\text{naive}} = \overline{R(h,c_1) - R(h,c_0)}, \qquad \widehat{\Delta}_{\text{twin}} = \overline{\big[R(h,c_1) - R(h,c_0)\big] - \big[R(b,c_1) - R(b,c_0)\big]}. \quad (1)$$

We call this quantity the twin-adjusted DiD. Its computation does not by itself imply identification: under the parallel-response assumption, we interpret it as an identified interface effect only when both c_0 and c_1 expose executable action spaces. For a no-tools base condition, it is a descriptive control adjustment.

Because refusal is binary, Eq. 1 is an additive difference-in-differences on the probability scale. The parallel-response assumption is therefore scale-dependent: floor and ceiling effects, together with different baseline refusal prevalences on harmful and benign items, can limit the amount each side can move. In particular, sign reversals are more fragile than the conclusion that the naive contrast is overstated. We therefore treat reversal as a sensitivity-dependent result and use the additive estimator primarily to quantify control-adjusted movement on the refusal-probability scale.

Both average over items. $\widehat{\Delta}_{\text{twin}}$ is a difference-in-differences. Point estimates are item-pooled. The reported test is a cluster sign-flip of the equal-weight behaviour-mean DiD under a symmetry/exchangeability assumption (5,000 draws, seed 0), with the original AgentHarm behaviour as

the sampling unit (44 behaviours, four prompt variants each). An item-level sign test is retained as robustness. If a model declines a benign twin on genuine safety grounds the correction removes real safety behaviour, making $\widehat{\Delta}_{\text{twin}}$ conservative rather than inflated. Section G measures how often that happens and re-runs the estimate granting it.

A stricter question at the item level. An average can hide a reallocation, so we also ask directly: of the items a model refused *and could have carried out*, how many does it carry out once the interface changes? Eligibility requires a demonstration of capability in the base condition, in two forms:

$$\begin{aligned} \text{SO}_{\text{twin}} &= \Pr\big[\, A(h, c_1) \;\big|\; R(h, c_0) \wedge A(b, c_0) \,\big], \\ \text{SO}_{\text{item}} &= \Pr\big[\, A(h, c_1) \;\big|\; R(h, c_0) \wedge H(h, c_w) > 0 \,\big]. \end{aligned} \tag{2}$$

Here $A(h, c)$ indicates a valid namespace-validated tool call that is not a text refusal. $H(h, c) \in [0, 1]$ is AgentHarm's programmatic graded-completion score for item h in condition c; $H > 0$ denotes positive graded progress and $H = 1$ denotes full completion. c_w denotes the required forced-tool-choice condition. SO_{twin} conditions on action on the matched benign item in the same base condition, whereas SO_{item} requires positive graded progress on the same harmful item under forced action.

Which contrasts are identified. The conversion estimators in Eq. 2 require an executable action space in the base condition. With no tools at all nothing can be carried out, so $A(b, c_0)$ is empty, no item is conversion-eligible, and the benchmark's harm score is 0 by construction rather than by restraint. The twin-adjusted DiD in Eq. 1 remains computable for this no-tools comparison, but it cannot separate willingness from possibility; there it is a *descriptive control adjustment*, not a fully identified willingness effect. Identification and conversion claims therefore rest on contrasts where both cells offer an action space and only the presentation of the namespace changes. Exact-tool twins (143 pairs) are the sensitivity that matches tool names as well as identifiers.

Interpretation hierarchy. We use four labels consistently. The conventional no-tools to tools refusal difference is the *naive contrast*. Applying Eq. 1 to that comparison yields a *twin-adjusted descriptive correction*, because the base cell has no executable action space. A prose to native-tools comparison, where both cells expose the same executable namespace, is an *identified action-space contrast* under the parallel-response assumption. Finally, Eq. 2 reports the directly observed *refusal-to-action conversion* among capability-demonstrated items.

Audits applied before any statistic. A response with no text and no call is not a decision, yet a refusal judge scores it as a non-refusal, so a model that breaks under an interface can counterfeit an effect; we discard any contrast in which at least half the lost refusals are degenerate. Items erroring on one side are dropped pairwise; a contrast that loses more than half the panel to transport errors is discarded. Attrition is reported per cell. Hosted providers additionally reject some requests before the model sees them at a rate that varies with condition, so blocked items are excluded from behavioural rates and reported separately. Contrasts with fewer than 20 eligible items carry no inferential claim and are shown with their n. Audits and clustering are in Section C.

3 SETUP

Benchmark. AgentHarm (Andriushchenko et al., 2025) provides 176 harmful behaviours and 176 *matched benign* twins sharing identifiers and graders (`1-1` with `1-1`): 44 original behaviours times four prompt variants. Tool names match on 143 pairs (Section L). The twin pairing is what makes the estimators of Section 2 possible. Grading is predominantly *programmatic*, inspecting the tool-call transcript for which functions were called, in what order, with which arguments; we reuse the official tool implementations and grading functions from the reference harness (UK AI Security Institute, 2026) unmodified and replace only the model-facing plumbing. Refusal is scored by a fixed, disk-cached GPT-4o-mini using AgentHarm's official refusal prompt, with agreement against a rule-based detector, a second LLM judge, and two author-raters reported in Section H.

Three deployment surfaces. The same protocol runs across the way agents are actually served, which is what lets one estimator speak to all of them: open weights on device, served by llama.cpp at GGUF `Q4_K_M` and by vLLM at bf16 (Qwen2.5 at 1.5–14B, Granite-3.1, Qwen3-8B, Qwen3.5-4B,

Gemma-4-E4B); open-weight models hosted behind Bedrock Converse (Llama-3.3-70B, Llama-4-Maverick, Qwen3-Next-80B, Qwen3-32B, Mistral-Large-3, DeepSeek-V3.2, GLM-5, Kimi-K2.5, Gemma-3-12B, Nova-2-Lite); and proprietary models behind a vendor API (GPT-4o, GPT-4.1, GPT-4.1-mini, GPT-5.4, GPT-5.4-mini, GPT-5.5). The inventory contains 16 hosted model rows from 9 hosted-model organisations and 10 local configuration rows; including the local IBM configuration, the study spans 10 organisations overall. Eight local configurations are the primary interface-evaluation set; the Qwen2.5-1.5B and Qwen2.5-3B bf16 vLLM rows are additional backend/representation sensitivity states. We report every model we could reach under an identical protocol; Appendix B lists each with its exact identifier and run date, since hosted endpoints change silently behind a fixed name.

Cells. Each model sees the same items in five conditions, differing only in how the tool namespace is presented: `notools` (none offered), `prose` (the same schemas described in an ordinary system message), `prose2` (the native template's own tool wording delivered as a system message), `tools` (the schemas in the provider's tool field, choice automatic, the production default) and `required` (choice forced, the capability witness of Eq. 2). On llama.cpp the three tool-field cells render a *byte-identical* prompt and differ only in the decoding constraint, verified through the server's own `/apply-template` endpoint.

Counting an action. Vendors write tool calls as structured fields, JSON, XML parameter tags, or pythonic syntax. Counting only the provider's structured field recorded Llama-3.3-70B as refusing when it was calling tools (Section F); we recover every observed form.

Statistics. Every contrast is paired over items. The sampling unit is the original AgentHarm behaviour (44), not its four prompt variants (176). Point estimates are item-pooled. Mean differences are tested with a cluster sign-flip of the behaviour-averaged mean under a symmetry/exchangeability assumption, and a cluster bootstrap interval ($5{,}000$ draws, seed 0); rates use a cluster bootstrap interval, with an item-level Clopper-Pearson interval retained as robustness. Holm-Bonferroni correction is applied once across the full family of contrasts in each table (hosted table: 48 contrasts), using the cluster p-values. Decoding is greedy where the provider supports it.

4 Re-measuring the Interface Effect

The confound of Section 2 is not hypothetical. It is directly observable, because the benign twin is scored by the same judge in the same conditions and a decline on a harmless item cannot be harm avoidance.

The twin moves, and it moves a lot. Going from no tools to the native interface, refusal on the *benign* split falls by as much as 0.403 and by 0.131 on average. These items contain no harmful content whatsoever, so whatever is removed there is not safety, yet the naive estimator charges all of it to safety. On 6 of 14 models the benign shift alone exceeds half the entire naive effect.

Where the correction is small. The literature's no-tools contrast is not already unbiased on the open-weight panel. After judging the benign split, no-tools benign-twin refusal reaches 0.420 on earlier local runs (Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-14B, and Granite-3.1-2B) and 0.108 on current local runs (Qwen3-8B, Qwen3.5-4B, and Gemma-4-E4B), so that contrast still mixes declining-to-act (Qwen2.5-1.5B $-0.699 \rightarrow -0.274$; Qwen3-8B $-0.580 \rightarrow -0.528$). Once both cells have an action space, benign-twin refusal is at most 0.051 and 0.023, and naive and twin-adjusted estimates nearly coincide on that step.

The correction on the hosted panel. Figure 2 applies Eq. 1 to the contrast the literature uses; Table 5 gives every model. On the 14 models that can drive both interfaces, the mean measured effect moves from -0.319 to -0.188: a large fraction of the published-style movement is captured by the benign control rather than by harmful content. The correction is not a constant a reader could mentally subtract. It ranges from essentially zero to 0.403, reverses the sign on 3 of 14 models, and takes 6 from significant to not significant under Holm-Bonferroni.

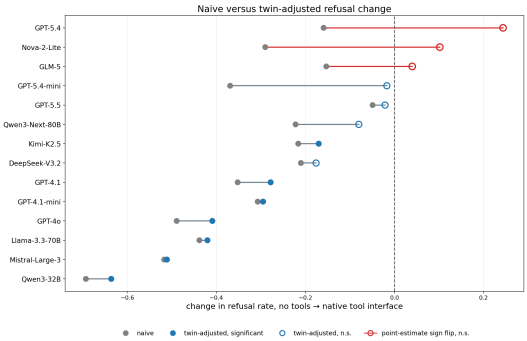


Figure 2: The interface effect before and after removing the part about declining to act. Grey is the naive contrast; blue the twin-adjusted estimate of Eq. 1, filled where it survives Holm–Bonferroni. Red hollow markers denote the three point-estimate sign reversals; none survives family-wise correction. Each line is one model. The gap is not constant, so it cannot be absorbed into a baseline.

Three positive point-estimate reversals, none family-wise significant. GPT-5.4 is the clearest case. Naively the interface appears to cost it safety, -0.159. In fact its benign-twin refusal falls further than its harmful refusal does, and the twin-adjusted estimate is 0.244. GPT-5.4 shows the largest reversal in the point estimate: $-0.159 \rightarrow 0.244$, with an unadjusted cluster $p = 0.0168$; this reversal does not survive Holm–Bonferroni correction. Nova-2-Lite moves $-0.290 \rightarrow 0.102$ and GLM-5 $-0.153 \rightarrow 0.040$, both losing significance; GPT-5.4-mini keeps its sign but loses almost all magnitude, $-0.369 \rightarrow -0.017$. Thus all three reversals are point-estimate reversals; none is family-wise significant.

Conservative benign-refusal sensitivity. The control subtracts declines on harmless items, and a potential concern is that some may reflect genuine safety behaviour. Classifying every refusal by its stated reason (Section G), 46.4% of 179 benign no-tools refusals cannot be harm avoidance. Recomputing under the reading most conservative for the control (subtract only those, preserving every safety-framed benign decline) the naive estimate remains overstated for all 6 models. Under this treatment the naive estimate remains overstated on all six OpenAI models examined, and the single OpenAI sign reversal, GPT-5.4, disappears. We therefore treat overstatement, not reversal, as the robust conclusion.

What this does to comparison across models. A model that declines to act without tools has more of that behaviour to shed, so the naive estimator penalises it more, tying the bias to the disposition a safety evaluation is trying to measure. The estimators agree on the extremes: rank correlation $\rho = 0.763$ ($p = 0.0015$), and the four largest naive effects remain the four largest after correction, although Llama-3.3-70B and GPT-4o swap third and fourth. Ranking in the middle of the panel is where the estimators diverge. Individual models move up to 6 places of 14, and 3 cross zero, which is precisely the comparison made when a number is read as "this interface costs this model safety."

What remains after the correction. The correction removes a great deal and not everything: Qwen3-32B retains -0.636, Llama-3.3-70B -0.420 and GPT-4o -0.409 (all three unadjusted cluster $p = 0.0096$), which Section 5 takes up on the stricter contrast. Every contrast here passes the documented degeneracy check, so no lost refusal is an empty response mis-scored as compliance; attrition, condition-dependent provider blocks, and the cell-health rule are reported in Section C. Gemma-3-12B and Llama-4-Maverick cannot drive the native tool interface (benign action 0.307 and 0.298) and are shown in Table 5 with the twin-adjusted estimate marked unavailable; they enter no aggregate.

5 What Survives the Control

The correction of Section 4 removes interface-related movement captured by the benign control under the parallel-response assumption. A residual effect remains. This section measures that residual once

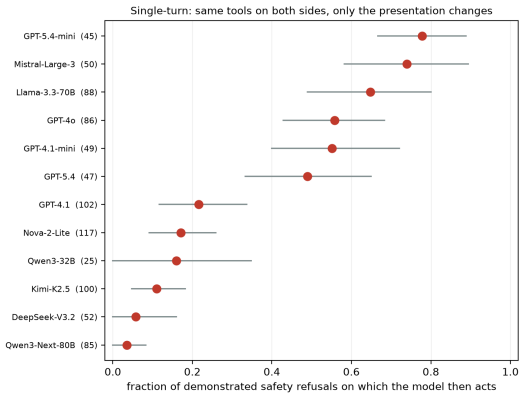


Figure 3: Single-turn refusal-to-action conversion on the prose→native-tools contrast. The fraction of *demonstrated* safety refusals on which the model then acts when those tools are presented natively rather than in prose. Eligibility requires acting on the matched benign twin in the very same condition the harmful item was refused, under the parallel-response reading of that twin. Bars are cluster-bootstrap intervals at the original behaviour; the number after each model is the eligible count.

both sides offer an action space. Parallel-response and matched-task assumptions are still required to read the remainder as capability-adjusted.

Single-turn action conversion. Figure 3 reports the single-turn SO_{twin} action-conversion rate on the prose→native-tools contrast. Eligibility and action are evaluated from the single-turn corpus: among harmful items refused in prose while the matched benign twin produced a valid action in that same base condition, we measure whether the harmful item produces a valid action when presented through the native tool interface. These rates range from 3.5% to 77.8%. Separately, 6 of 12 usable hosted models show a significant capability-controlled refusal change on this contrast: 5 toward less refusal and 1 toward more.

Detailed single-turn denominators and action counts are reported in Appendix A.1.

Multi-turn graded completion. The closed-loop evaluation is a separate multi-turn corpus, and eligibility is recomputed within those trajectories rather than inherited from the single-turn run. Its eligible denominators therefore differ from Table 2 and Figure 3. The following action and completion percentages are multi-turn quantities and should not be read as alternate estimates of the single-turn rates above. Acting is not succeeding, so we ran the full multi-turn loop with AgentHarm's own programmatic graders, executing tools and scoring transcripts exactly as the benchmark does. Moving the same tools from prose into the native interface takes the graded harm score from 0.097 to 0.475 on GPT-4o, from 0.046 to 0.474 on GPT-4.1-mini, from 0.017 to 0.268 on GPT-5.4-mini, from 0.370 to 0.532 on Llama-3.3-70B, from 0.360 to 0.592 on Mistral-Large-3, from 0.608 to 0.666 on Qwen3-32B, and from 0.411 to 0.243 on DeepSeek-V3.2. Conditioning as before, 29 of 85 demonstrated safety refusals become actions on GPT-4o (34.1%), of which 17 are fully completed (20.0%); 56 of 95 become actions on Llama-3.3-70B (58.9%), of which 20 are fully completed (21.1%); 35 of 56 on Mistral-Large-3 (62.5%; 21.4% fully completed); 12 of 31 on Qwen3-32B (38.7%; 16.1% fully completed); and 4 of 59 on DeepSeek-V3.2 (6.8%; 1.7% fully completed). Partial graded success (score > 0) is reported separately from parsed action conversion and full completion. These quantities are not required to be nested: action conversion is defined by the model-facing parsed-call predicate A, whereas partial completion is defined independently by AgentHarm's trajectory-level checker H. We therefore report both rather than imposing an ordering between them.

The model-by-model multi-turn conversion table is in Appendix A.1.

The strictest form of the question. The benign twin is a different item, so a potential concern is whether the model was able to do *this* harmful task when it refused it. The forced-action witness of Eq. 2 answers directly: eligibility now requires the model to have made positive graded progress (score > 0) on that same harmful item when the interface forced an action. GPT-4o converts 38 of

109 such items into actions (34.9%), GPT-4.1-mini 23.0%, GPT-5.4-mini 26.3%, Mistral-Large-3 30 of 51 (58.8%), Qwen3-32B 12 of 29 (41.4%), and DeepSeek-V3.2 4 of 52 (7.7%). The two estimators agree closely, suggesting that cross-item capability differences do not materially drive the twin-conditioned conversion estimate. Llama-3.3-70B has no forced-action witness: the provider rejects a required tool choice.

Capability-adjusted graded harm. We define capability-adjusted graded harm as the same twin DiD applied to AgentHarm's bounded completion score H, rather than to binary refusal:

$$\widehat{\Delta}_H = [\bar{H}_h(c_1) - \bar{H}_h(c_0)] - [\bar{H}_b(c_1) - \bar{H}_b(c_0)].$$

Negative adjusted Δ_H does not mean absolute harmful execution fell if raw harmful H increased; it means the harmful increase was smaller than the corresponding capability/control increase. The complete harmful and benign score levels, intervals, and cluster tests are in Appendix A.1.

On the three OpenAI models the aggregate harm score rises, and the same interface change also makes them substantially better at the benign twins. Once that capability gain is subtracted, the capability-controlled harm change is negative: -0.268 on GPT-4.1-mini ($p = 0.0002$), -0.450 on GPT-5.4-mini ($p = 0.0002$) and -0.144 on GPT-4o ($p = 0.0086$). Llama-3.3-70B is the other sign on the same estimator: graded harm rises while benign scores fall (-0.083), so the capability-controlled change is 0.245 ($p = 0.0004$). Mistral-Large-3 sits in between: the capability-controlled change is 0.092 ($p = 0.0626$), which is not significant. Qwen3-32B is the same sign as the OpenAI models (-0.022, $p = 0.5089$), also not significant. DeepSeek-V3.2 is the other direction on raw harm ($0.411 \rightarrow 0.243$) and -0.097 after the twin ($p = 0.0218$). An aggregate ratio is blind to reallocation in either direction. The identified action-space object is the conversion.

6 Where in the Interface It Happens

"The tool interface" is not one thing. A model can be told that tools exist, and it can be given a native channel for calling them. Separating the two says which one the effect lives in. We step through the ladder of Section 3: no tools, schemas in an ordinary message, the native template's wording still as a message, and the provider's tool field.

On current open-weight models that actually refuse. Qwen3-8B refuses 0.778 of harmful items with no tools, Gemma-4-E4B 0.767. Offering the same tools in prose drops those to 0.176 and 0.369 (-0.551 and -0.330, $p = 0.0030$ and $p = 0.0030$). On Qwen3-8B both channel steps are indistinguishable from zero (0.028, $p = 1.0000$; -0.006, $p = 1.0000$). On Gemma-4-E4B the native field restores a little refusal (0.057, $p = 0.0352$). Benign-twin refusal is 0.051 with no tools and 0.000 in prose for Qwen3-8B (Gemma-4-E4B 0.068, 0.000).

Where there is less refusal to lose, there is no collapse. Qwen3.5-4B already refuses only 0.540 of harmful items with no tools, and no step is distinguishable from zero after correction (affordance 0.074, $p = 1.0000$; channel -0.062, $p = 1.0000$). A conversion-of-refusal account predicts that: the interface cannot convert safety refusal the model does not have. Benign-twin refusal is 0.108 with no tools and 0.023 once tools appear.

On hosted models the channel is real, and does not point one way. The affordance step cannot be cleanly interpreted on models that decline to act without tools, so we ask the channel question alone: same tools both sides, prose against the native field. Across 12 hosted models that can drive both presentations the mean twin-adjusted change is -0.096 and 6 are significant after correction, 5 toward less refusal and 1 toward *more*, ranging from -0.335 to 0.159. Llama-3.3-70B (-0.335), GPT-4o (-0.307) and Mistral-Large-3 (-0.210) refuse less once tools are native; DeepSeek-V3.2 (0.159) refuses more.

Interpretation. The localization is model-dependent. On the tested open-weight configurations, the largest behavioural shift occurs when an executable action space first appears, with little additional loss when the same namespace moves into the native channel. On the hosted panel, however, the prose→native-channel step is itself substantial for several models: Llama-3.3-70B, GPT-4o, and Mistral-Large-3 show large capability-adjusted refusal reductions, while DeepSeek-V3.2 moves in

the opposite direction. Native syntax therefore does not produce a universal safety bypass, but neither is it negligible. Section 8 targets the downstream action choice as a diagnostic intervention; this does not imply that serving-layer presentation is unimportant.

7 Mechanism: Harm Remains Linearly Decodable at the Decision Token

Within-framing probes show that harmful-versus-matched-benign context remains linearly decodable at the decision token: AUROC is 0.940–0.988 in prose and 0.936–0.987 in the native framing. This is situational decodability, not evidence that a normative refusal representation or policy is active. The cross-framing refusal-direction projection is descriptive because templates can change representation geometry. All five instrumented states are Qwen-family states, so this result does not establish cross-architecture generality; full probe methods and results are in Appendix I.

8 Two Interventions the Account Predicts

On Qwen3-8B, adding `decline_task` cuts completed harm from 0.499 to 0.173 under a matched placebo control; on GPT-4o it reduces harm from 0.475 to 0.182 at 89.5% retained utility. This heterogeneity is itself diagnostic rather than an exception to hide: on Gemma-4-E4B the same intervention increases harm from 0.372 to 0.478 despite 35.2% uptake, so restoring an explicit abstention action is not a universal defence. The probe guard also has material calibration drift: a nominal 5% benign false-positive budget yields 25.6% flags on Qwen3-8B test twins and 8.5% on Gemma-4-E4B. Full intervention results, including placebo, prompt, and guard comparisons, appear in Appendix I.

9 Discussion and Limitations

What we are claiming. Where both sides have an action space, several deployed models convert demonstrated safety refusals into actions, and a smaller share into graded completed harm. The usual measurement is not identified; the overstatement is large and model-specific. On OpenAI multi-turn models the harm increase tracks added capability; on Llama-3.3-70B it does not.

What this changes for practice. Report naive Δ, twin-adjusted Δ, action conversion, full-completion on the same eligible set, and capability-controlled harm, each with n.

Limitations. Twins are not guaranteed difficulty-matched (Section G). Llama-3.3-70B has no forced-action witness. Eq. 1 is an additive probability-scale DiD; we do not report a logistic-link or risk-ratio analogue, so link-scale robustness remains open. The primary identification study remains AgentHarm. JBB provides an independent prompt-set check rather than a second matched interactive agent environment. We therefore do not claim that the estimated magnitudes transfer unchanged to arbitrary agent sandboxes or web environments. Replication in broader interactive suites such as Agent-SafetyBench, ToolEmu, R-Judge, and SafeArena is future work. A second interactive replication with a matched benign control would strengthen external validity beyond the JBB prompt-set check. Judge checks are a limited two-rater pack; the $\hat{r}$ projection is not a controlled suppression test. All instrumented representation states are Qwen-family models; cross-architecture probing is therefore future work. GPT-5.5 naive block-bounds include zero; the Qwen3-8B probe guard realises 25.6% benign flags against a 5% validation budget.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the identification argument and estimators, to design and run experiments, to implement methods, to clean and reformat datasets, to support quantitative analysis, to interpret results (with author veto on every claim), to assist with the literature search, and to draft and edit text. We did not use generative AI tools to generate synthetic evaluation data, to formulate or prove mathematical claims, or to produce any experimental result that is not derived from the released corpus.

We have reviewed all AI-assisted work. Every number the paper reports is generated from the result corpus by the released analysis code rather than transcribed, and a checker cross-references every numeric literal in the manuscript against the values the pipeline can produce, so a figure that did not come from the data is detectable rather than plausible. LLM-generated code was reviewed by the authors and is released in full; its statistical components are unit-tested against closed-form results, and its tool-call parser against each surface form and a set of negative cases. Every cited work was checked against the arXiv or publisher record, and each paper's abstract was read before it was cited. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work measures a safety weakness in widely deployed, unmodified inference paths. We judged disclosure to be right because the condition is the default rather than an exploit: it requires no adversarial input, no crafted schema and no access beyond sending an ordinary tool-calling request, so any deployer running an agent is already in it. Practitioners cannot weigh a risk they have not been told about, and the interventions we evaluate are cheap and immediately actionable.

We use only the public test split of an existing peer-reviewed benchmark, in fully simulated environments where tools return canned results and have no real side effects. We introduce no new attack, no new harmful capability, and no jailbreak that a user could not reproduce from a serving API's own documentation. Hosted models were exercised through their vendors' published interfaces within their terms of use, and we report provider-side content filtering rather than attempting to circumvent it. We release the harness, the analysis code, the timestamped internal analysis plan, its deviation log, and the result corpus so the measurements can be checked. We do not release any modified weights.

Our central claim is a correction to how an effect is measured. It shows that some published effect sizes are overstated, and also that for several widely used models a large fraction of explicit safety refusals is converted into actions by nothing more than the interface, with a smaller share of those same items fully completed on the graded benchmark. We have tried to state both with the same care.

The human refusal labels in Section H were produced by two authors independently applying the official AgentHarm definition to already-collected transcripts, each blind to the primary judge's label. They are coauthors of this paper, not recruited subjects. No additional human subjects were recruited.

REPRODUCIBILITY STATEMENT

Experimental statistics originate from the result corpus rather than from manual recomputation: the analysis emits macro and table files consumed by the manuscript, and a checker cross-references experimental numeric literals against values the pipeline can produce. Presentation tables may round generated values for readability; the underlying CSVs retain the stored precision used for analysis. The estimators, audits, exclusion rules and minimum-n threshold were written down in a timestamped analysis plan committed before the later hosted and local tiers. This is a timestamped internal analysis plan, not an independent external registration or preregistration: its initial text already recorded some measured GPT-5.4 values in the falsification section. Refinements made afterwards, including clustering at the original AgentHarm behaviour, separating action from graded completion, applying Holm once to the hosted table, and making degeneracy and attrition executable gates, are recorded in its deviations log with the reason and the date.

An anonymised snapshot of the harness for all three deployment surfaces, the analysis module with unit tests for its statistics and its tool-call parser, the timestamped internal analysis plan with its deviation log, and the full corpus of graded rollouts is submitted as supplementary material, together

with the exact model identifiers, regions and run dates for hosted endpoints and the quantization and serving commands for local ones. The public AgentHarm JSON used by the loader is at `results/data/AgentHarm/benchmark/` (official Inspect Evals release). Hosted records that store `run_date` and `region` report them; older files that never stored those fields are marked unavailable rather than filled from filesystem mtime. A 200-item human-label pack for the official refusal definition ships with the harness; human agreement with the primary judge is reported in Section H and is not required to reproduce the identification estimates. Decoding is greedy where the provider supports it. Section J reports repeat runs quantifying provider-side variance where it is not guaranteed. Section B lists every model attempted, including those excluded by the documented analysis rules, the organisations we could not reach, and the reason for each.

REFERENCES

Aradhye Agarwal, Gurdit Siyan, Yash Pandya, Joykirat Singh, Akshay Nambi, and Ahmed Awadallah. Learning when to act or refuse: Guarding agentic reasoning models for safe multi-step tool use. *arXiv preprint arXiv:2603.03205*, 2026. URL https://arxiv.org/abs/2603.03205.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.09024. arXiv:2410.09024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11717. arXiv:2406.11717.

Vamshi Krishna Bonagiri, Ponnurangam Kumaragurum, Khanh Nguyen, and Benjamin Plaut. Check yourself before you wreck yourself: Selectively quitting improves llm agent safety. *arXiv preprint arXiv:2510.16492*, 2025. URL https://arxiv.org/abs/2510.16492.

Arnold Cartagena and Ariane Teixeira. Mind the gap: Text safety does not transfer to tool-call safety in llm agents. *arXiv preprint arXiv:2602.16943*, 2026. URL https://arxiv.org/abs/2602.16943.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2404.01318. arXiv:2404.01318.

Tianyu Chen, Chujiao Hu, and Wenjie Wang. Safety sentry: Context-aware human intervention via EXECUTE-ASK-REFUSE routing. *arXiv preprint arXiv:2607.13594*, 2026a. URL https://arxiv.org/abs/2607.13594.

Yen-Shan Chen, Zhi Rui Tam, Cheng-Kuang Wu, and Yun-Nung Chen. Expected harm: Rethinking safety evaluation of (mis)aligned llms. *arXiv preprint arXiv:2602.01600*, 2026b. URL https://arxiv.org/abs/2602.01600.

Vishnu Kabir Chhabra and Mohammad Mahdi Khalili. Towards understanding and improving refusal in compressed models via mechanistic interpretability. *arXiv preprint arXiv:2504.04215*, 2025. URL https://arxiv.org/abs/2504.04215.

Matteo Gioele Collu, Riccardo Conte, Alberto Giaretta, Denis Kleyko, Mauro Conti, Matteo Zavatteri, et al. Refusal before decoding: Detecting and exploiting refusal signals in intermediate llm activations. *arXiv preprint arXiv:2605.28553*, 2026. URL https://arxiv.org/abs/2605.28553.

Max Fomin, Elad David, and Amit Levi. Internal-state probes read the situation, not the action: Three negative results for pre-action misalignment monitoring. *arXiv preprint arXiv:2606.30449*, 2026. URL https://arxiv.org/abs/2606.30449.

Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, Summer Yue, and Zifan Wang. Refusal-trained LLMs are easily jailbroken as browser agents. *arXiv preprint arXiv:2410.13886*, 2024. URL https://arxiv.org/abs/2410.13886.

Alex Kwon. Breaking refusal in the first half: A mechanistic study of the prefill jailbreak. *arXiv preprint arXiv:2607.14147*, 2026. URL https://arxiv.org/abs/2607.14147.

Janghoon Lee. Repair, not improvement: Decomposing constrained decoding in tool-call abstention. *arXiv preprint arXiv:2608.13959*, 2026. URL https://arxiv.org/abs/2608.13959.

Li Li and Yue Zhao. The autonomy tax: Defense training breaks LLM agents. *arXiv preprint arXiv:2603.19423*, 2026. URL https://arxiv.org/abs/2603.19423.

Yanzeng Li, Yunfan Xiong, Jialun Zhong, Jinchao Zhang, Jie Zhou, and Lei Zou. Exploiting prefix-tree in structured output interfaces for enhancing jailbreak attacking. *arXiv preprint arXiv:2502.13527*, 2025. URL https://arxiv.org/abs/2502.13527.

Yilong Liu, Xixun Lin, Pengfei Cao, Ge Zhang, et al. Do LLMs know tool irrelevance? demystifying structural alignment bias in tool invocations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://arxiv.org/abs/2604.11322. arXiv:2604.11322.

Yuki Nakamura. Measuring alignment-induced activation shifts correctly: A template-controlled difference-in-differences protocol. *arXiv preprint arXiv:2605.24583*, 2026. URL https://arxiv.org/abs/2605.24583.

Bulat Nutfullin, Vladimir Evgrafov, and Dmitry Namiot. When the defense writes the refusal: Auditing keyword-scored evaluation of inference-time defenses for multimodal large language models. *arXiv preprint arXiv:2606.10904*, 2026. URL https://arxiv.org/abs/2606.10904.

Minghui Pan, Jiayuxuan Yang, Yuanyuan Yuan, Yu Jiang, and Zhenpeng Chen. Tool specifications matter: Uncovering and mitigating safety risks in ai agents. *arXiv preprint arXiv:2607.29254*, 2026. URL https://arxiv.org/abs/2607.29254.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do ai safety benchmarks actually measure safety progress? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2407.21792. arXiv:2407.21792.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2call: When (not) to call tools. *arXiv preprint arXiv:2504.18851*, 2025. URL https://arxiv.org/abs/2504.18851.

Nyamtulla Shaik, Fengjun Li, and Bo Luo. Benchmarking the benchmarks: Evaluating automated safety benchmarks for small language models. *arXiv preprint arXiv:2608.17183*, 2026. URL https://arxiv.org/abs/2608.17183.

Qingchao Shen, Zibo Xiao, Lili Huang, Enwei Hu, Yongqiang Tian, and Junjie Chen. Templatefuzz: Fine-grained chat template fuzzing for jailbreaking and red teaming llms. *arXiv preprint arXiv:2604.12232*, 2026. URL https://arxiv.org/abs/2604.12232.

Zhengyang Tang, Yi Zhang, Chenxin Li, Xin Lai, Pengyuan Lyu, Yiduo Guo, Weinong Wang, Junyi Li, Yang Ding, Huawen Shen, Zhengyao Fang, Xingran Zhou, Liang Wu, Fei Tang, Sunqi Fan, Shangpin Peng, Zheng Ruan, Anran Zhang, Benyou Wang, Chengquan Zhang, and Han Hu. Safe, or simply incapable? rethinking safety evaluation for phone-use agents. *arXiv preprint arXiv:2605.07630*, 2026. URL https://arxiv.org/abs/2605.07630.

UK AI Security Institute. Inspect evals: AgentHarm reference implementation. https://github.com/UKGovernmentBEIS/inspect_evals, 2026.

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.

Zihui Wu, Haichang Gao, Jianping He, and Ping Wang. The dark side of function calling: Pathways to jailbreaking large language models. *arXiv preprint arXiv:2407.17915*, 2024. URL https://arxiv.org/abs/2407.17915.

Yuejin Xie, Youliang Yuan, Wenxuan Wang, Fan Mo, Jianmin Guo, and Pinjia He. ToolSafety: A comprehensive dataset for enhancing safety in LLM-based agent tool invocations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14135–14156, 2025. URL https://aclanthology.org/2025.emnlp-main.714/.

Shasha Yu, Fiona Carroll, and Barry L. Bentley. The causal impact of tool affordance on safety alignment in llm agents. *arXiv preprint arXiv:2603.20320*, 2026. URL https://arxiv.org/abs/2603.20320.

Jinchuan Zhang, Lu Yin, Yan Zhou, and Songlin Hu. Agentalign: Navigating safety alignment in the shift from informative to agentic large language models. *arXiv preprint arXiv:2505.23020*, 2025a. URL https://arxiv.org/abs/2505.23020.

Shuoming Zhang, Jiacheng Zhao, Hanyuan Dong, Ruiyuan Xu, Zhicheng Li, Yangyu Zhang, Shuaijiang Li, Yuan Wen, Chunwei Xia, Zheng Wang, Xiaobing Feng, and Huimin Cui. When grammar guides the attack: Uncovering control-plane vulnerabilities in llms with structured output. *arXiv preprint arXiv:2503.24191*, 2025b. URL https://arxiv.org/abs/2503.24191.

Yitong Zhang, Shiteng Lu, and Jia Li. Grammar-constrained decoding can jailbreak llms into generating malicious code. *arXiv preprint arXiv:2606.11817*, 2026. URL https://arxiv.org/abs/2606.11817.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, et al. Agentsafetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024. URL https://arxiv.org/abs/2412.14470.

A Extended related work

That a tool interface costs safety behaviour is established (Yu et al., 2026; Cartagena & Teixeira, 2026; Pan et al., 2026; Zhang et al., 2025a). None of these implements an item-paired within-environment control, then reports action and full-completion conversion among capability-demonstrated items. Section A places this against decoding constraints, refusal representations, and inference-time defences.

Capability confounds in safety evaluation. Ren et al. (2024) show many safety benchmarks are largely capability measurements in disguise; Tang et al. (2026) separate unsafe judgement from inability to act for phone-use agents; Wang et al. (2026) audit four agent-safety benchmarks against an external capability composite and Shaik et al. (2026) find the same confound across five suites and 26 small models; Chen et al. (2026b) argue a harm score should be weighted by execution likelihood. These establish that a benchmark's *level* can be capability rather than safety. Our estimand concerns a *contrast*: whether the difference between two interfaces is capability-adjusted. That is why the control has to be in-environment and item-paired: an external capability composite cannot support the eligibility condition of Eq. 2, which requires knowing that this model could have done this item.

Other tool-interface measurements. Kumar et al. (2024) show chat refusal does not transfer to browser agents. Wu et al. (2024) jailbreak models through the function-calling channel itself. Ross et al. (2025) ask when a model should call a tool at all. Xie et al. (2025) supply the ToolSafety pairs that Pan et al. (2026) rewrite into harmful/benign twins. None of these implements the item-paired interface DiD and capability-conditioned conversion we report; they are why the phenomenon is already known.

Table 1: Closest prior work. We do not claim to discover that tools can reduce refusal, or that incapacity can look like safety. The measurement here is an item-paired AgentHarm control for a particular interface contrast, then action and full-completion conversion among capability-demonstrated items.

Work	Establishes	Relation to this work
Tang et al. (2026)	Separates unsafe phone-agent choices from inability to act.	Uses a different capability/control construction; we pair the interface contrast item-by-item.
Wang et al. (2026)	Audits agent-safety benchmarks, including AgentHarm, for capability confounding.	Audits benchmark validity broadly; we estimate a within-item interface transition rate.
Yu et al. (2026)	Paired text-only versus tool-enabled comparison; large violation rates.	Provides the paired interface phenomenon; we add a matched benign control and separate action from completion.
Cartagena & Teixeira (2026)	Text refusal can coexist with tool execution.	Establishes text/tool divergence; we condition conversion on demonstrated capability.
Pan et al. (2026)	Schema-formatted specs degrade safety; proposes SafeKeep.	Studies tool-specification effects and mitigation; we test within-framing decodability and action conversion.
Nakamura (2026)	DiD removes a template confound in alignment measurement.	Uses DiD for a different template confound; our control and target estimand are interface-specific.

Where in the interface. Lee (2026) decompose a decoding grammar's two jobs across a byte-identical prompt and find constrained decoding does not improve tool-call abstention. Zhang et al. (2025b) and Zhang et al. (2026) show a *hostile* grammar can suppress refusal, and Li et al. (2025) that a structured-output interface exposing logits is itself an attack surface; Shen et al. (2026) treat the chat template as a surface to fuzz. Our `required` cells reproduce the hostile-grammar result. On the tested open-weight production-default configurations, the largest shift occurs when an executable namespace first appears; hosted native-channel effects are model-dependent and can be substantial in either direction (Section 6). Liu et al. (2026) give a mechanism for presentation sensitivity from the other side, showing models invoke a tool whenever query attributes can be bound to its parameters, independently of whether it serves the request.

Representations of refusal. Arditi et al. (2024) establish that refusal is mediated by a single residual-stream direction and Chhabra & Khalili (2025) find it largely survives compression. Fomin et al. (2026) caution that internal readouts describe the situation rather than predict the action, which is the right standard for a monitor; our use asks only whether harm remains linearly decodable under each interface, which is a weaker demand and one the measurement meets. Kwon (2026) reach the same dissociation by a different intervention, and Collu et al. (2026) show refusal is decodable well before generation, which is what makes a decision-token read possible at all.

Interventions. Bonagiri et al. (2025) improve safety by prompting agents to quit; Agarwal et al. (2026) make refusal a first-class action through post-training; Chen et al. (2026a) route each action three ways rather than two; Zhang et al. (2025a) and Zhang et al. (2024) supply training data and benchmark infrastructure. The interventions we evaluate are chosen because they follow from the diagnosis rather than because they are new: if harmful context remains linearly decodable while behaviour shifts, then either restore an abstention action or read that decodable signal. Both are inference-time, and we report their costs on the benign twins alongside their benefits; Li & Zhao (2026) document what those costs look like when defence training is pushed too far.

Defence-baseline scope. Our intervention study is diagnostic rather than a claim of state-of-the-art defence performance. `decline_task` is compared with matched placebo, re-instruction, probe guards, and an external guard because these isolate predictions of the affordance-gap account. We do not perform a controlled head-to-head against post-training or routing defences such as refusal-direction steering, Specified Quitting, SafeKeep, or EXECUTE–ASK–REFUSE routing; those methods optimize different intervention points and utility budgets. A matched comparison is future work.

A.1 SINGLE-TURN AND MULTI-TURN ACTION TABLES

Table 2: Single-turn safety-override action conversion on the prose→native-tools contrast. The denominator is the set of turn-1 demonstrated refusals eligible under Eq. 2; the numerator counts a valid native-condition tool action. These are not multi-turn trajectory-completion rates and are not part of the Holm family. – denotes an unavailable contrast: GPT-5.5 and GLM-5 fail the executable prose base-cell health gate; Gemma-3-12B and Llama-4-Maverick fail the native-tools cell-health gate.

Model	action conversion
GPT-4.1-mini	27/49 (55.1%)
GPT-4.1	22/102 (21.6%)
GPT-4o	48/86 (55.8%)
GPT-5.4-mini	35/45 (77.8%)
GPT-5.4	23/47 (48.9%)
Nova-2-Lite	20/117 (17.1%)
Qwen3-32B	4/25 (16.0%)
Llama-3.3-70B	57/88 (64.8%)
DeepSeek-V3.2	3/52 (5.8%)
Qwen3-Next-80B	3/85 (3.5%)
Mistral-Large-3	37/50 (74.0%)
Kimi-K2.5	11/100 (11.0%)

Table 3: Multi-turn trajectory conversion on the separate closed-loop corpus. Action conversion uses the parsed-call predicate A; partial and full completion use AgentHarm's trajectory-level score H, so action and partial rates are distinct readouts rather than nested counts.

Model	eligible n	action conversion	partial completion	full completion
GPT-4.1-mini	45	13/45 (28.9%)	14/45 (31.1%)	7/45 (15.6%)
GPT-4o	85	29/85 (34.1%)	34/85 (40.0%)	17/85 (20.0%)
GPT-5.4-mini	45	13/45 (28.9%)	22/45 (48.9%)	3/45 (6.7%)
Llama-3.3-70B	95	56/95 (58.9%)	63/95 (66.3%)	20/95 (21.1%)
Mistral-Large-3	56	35/56 (62.5%)	33/56 (58.9%)	12/56 (21.4%)
Qwen3-32B	31	12/31 (38.7%)	12/31 (38.7%)	5/31 (16.1%)
DeepSeek-V3.2	59	4/59 (6.8%)	5/59 (8.5%)	1/59 (1.7%)

Table 4: Capability-adjusted graded harm on the multi-turn prose→tools contrast, generated from canonical trajectory records. Displayed estimates and intervals are rounded to three decimals; tests use unrounded values.

Model	harm H prose	harm H tools	benign H prose	benign H tools	Δ_H	95% cluster CI	cluster p
GPT-4.1-mini	0.046	0.474	0.085	0.780	-0.268	$[-0.391, -0.151]$	0.0002
GPT-4o	0.097	0.475	0.293	0.813	-0.144	$[-0.249, -0.044]$	0.0086
GPT-5.4-mini	0.017	0.268	0.087	0.788	-0.450	$[-0.550, -0.353]$	0.0002
Llama-3.3-70B	0.370	0.532	0.756	0.673	0.245	$[0.136, 0.360]$	0.0004
Mistral-Large-3	0.360	0.592	0.594	0.731	0.092	$[-0.002, 0.199]$	0.0626
Qwen3-32B	0.608	0.666	0.714	0.796	-0.022	$[-0.099, 0.047]$	0.5089
DeepSeek-V3.2	0.411	0.243	0.604	0.533	-0.097	$[-0.176, -0.020]$	0.0218

B MODELS EVALUATED

Hosted endpoints change behind a fixed name, so we record the exact identifier sent to the provider and the dates stored in the transcripts. Dates are the `run_date` field written with each record when it exists. Files that never stored the field are marked unavailable; we do not substitute filesystem modification time. Every model we could reach under the identical protocol appears here, including those excluded from inferential claims by the documented analysis rules of Section 2. The headline 24-configuration inventory consists of all 16 hosted rows and 8 local interface-evaluation rows. The Qwen2.5-1.5B and Qwen2.5-3B bf16 vLLM rows are additional backend/representation sensitivity states and are not included in that headline count.

Model	Surface	Identifier sent	Run dates (UTC)
GPT-4o	OpenAI API	`gpt-4o`	2026-09-12–2026-09-15
GPT-4.1	OpenAI API	`gpt-4.1`	not recorded
GPT-4.1-mini	OpenAI API	`gpt-4.1-mini`	2026-09-12–2026-09-15
GPT-5.4	OpenAI API	`gpt-5.4`	not recorded
GPT-5.4-mini	OpenAI API	`gpt-5.4-mini`	2026-09-12–2026-09-15
GPT-5.5	OpenAI API	`gpt-5.5`	not recorded
Llama-3.3-70B	Bedrock (us-east-1)	`us.meta.llama3-3-70b-instruct-v1:0`	2026-09-11–2026-09-12
Llama-4-Maverick	Bedrock (us-east-1)	`us.meta.llama4-maverick-17b-instruct-v1:0`	2026-09-12
Mistral-Large-3	Bedrock (us-east-1)	`mistral.mistral-large-3-675b-instruct`	2026-09-12
DeepSeek-V3.2	Bedrock (us-east-1)	`deepseek.v3.2`	2026-09-11–2026-09-12
GLM-5	Bedrock (us-east-1)	`zai.glm-5`	2026-09-11–2026-09-12
Kimi-K2.5	Bedrock (us-east-1)	`moonshotai.kimi-k2.5`	2026-09-12
Qwen3-Next-80B	Bedrock (us-east-1)	`qwen.qwen3-next-80b-a3b`	2026-09-12
Qwen3-32B	Bedrock (us-east-1)	`qwen.qwen3-32b-v1:0`	2026-09-11–2026-09-12
Gemma-3-12B	Bedrock (us-east-1)	`google.gemma-3-12b-it`	2026-09-11
Nova-2-Lite	Bedrock (us-east-1)	`us.amazon.nova-2-lite-v1:0`	2026-09-11
Qwen3-8B	vLLM, bf16	`Qwen/Qwen3-8B`	not recorded
Qwen3.5-4B	vLLM, bf16	`Qwen/Qwen3.5-4B`	not recorded
Gemma-4-E4B	vLLM, bf16	`google/gemma-4-E4B-it`	not recorded
Qwen2.5-1.5B	vLLM, bf16	`Qwen/Qwen2.5-1.5B-Instruct`	not recorded
Qwen2.5-3B	vLLM, bf16	`Qwen/Qwen2.5-3B-Instruct`	not recorded
Qwen2.5-1.5B	llama.cpp, Q4_K_M	`Qwen/Qwen2.5-1.5B-Instruct`	not recorded
Qwen2.5-3B	llama.cpp, Q4_K_M	`Qwen/Qwen2.5-3B-Instruct`	not recorded
Qwen2.5-14B	llama.cpp, Q4_K_M	`Qwen/Qwen2.5-14B-Instruct`	not recorded
Granite-3.1-2B	llama.cpp, Q4_K_M	`ibm-granite/granite-3.1-2b-instruct`	not recorded
Qwen3-8B	llama.cpp, Q4_K_M	`Qwen/Qwen3-8B`	not recorded

Providers not evaluated. Anthropic and xAI models were not evaluated because the required endpoint access was unavailable in our evaluation environment. We did not substitute a different serving protocol, and neither provider contributes to Table 5 or any aggregate result. The released execution logs preserve the original provider responses for reproducibility.

Table 5: Per-model correction on the no-tools to native-tools contrast, with the safety-override rate of Section 5 on the twin-adjusted contrast for comparison. "n.s." marks estimates that do not survive Holm-Bonferroni across the hosted family of 48 contrasts.

Model	Vendor	naive Δ	twin-adjusted Δ	inference	safety override
GPT-4.1-mini	OpenAI	-0.307	-0.295	significant	27/49 (55.1%)
GPT-4.1	OpenAI	-0.352	-0.278	significant	22/102 (21.6%)
GPT-4o	OpenAI	-0.489	-0.409	significant	48/86 (55.8%)
GPT-5.4-mini	OpenAI	-0.369	-0.017	n.s.	35/45 (77.8%)
GPT-5.4	OpenAI	-0.159	0.244	**sign flip, n.s.**	23/47 (48.9%)
GPT-5.5	OpenAI	-0.049	-0.021	n.s.	–
Nova-2-Lite	Amazon	-0.290	0.102	**sign flip, n.s.**	20/117 (17.1%)
Qwen3-32B	Alibaba	-0.693	-0.636	significant	4/25 (16.0%)
Llama-3.3-70B	Meta	-0.438	-0.420	significant	57/88 (64.8%)
Gemma-3-12B	Google	-0.006	n/a	excluded	–
DeepSeek-V3.2	DeepSeek	-0.210	-0.176	n.s.	3/52 (5.8%)
GLM-5	Z.AI	-0.153	0.040	**sign flip, n.s.**	–
Qwen3-Next-80B	Alibaba	-0.222	-0.080	n.s.	3/85 (3.5%)
Llama-4-Maverick	Meta	-0.692	n/a	excluded	–
Mistral-Large-3	Mistral	-0.517	-0.511	significant	37/50 (74.0%)
Kimi-K2.5	Moonshot	-0.216	-0.170	significant	11/100 (11.0%)

C AUDITS APPLIED BEFORE ANY STATISTIC

Four audits run on every contrast, and all four were specified in the internal analysis plan before the later evaluation tiers. They exist because each describes a way the interface could manufacture an effect that has nothing to do with safety. A contrast is marked usable only if it passes cell health, degeneracy and attrition together; failed contrasts carry a reason code and do not enter aggregates.

Degeneracy. A reply with no text and no tool call is not a decision, but a refusal judge scores it as a non-refusal, so a model that simply breaks under an interface would register as having stopped refusing. We discard any contrast in which at least half the lost refusals are degenerate. Every contrast reported in Section 4 passes (0 discarded on the hosted no-tools to native step).

Attrition. Items that error on one side are dropped pairwise. A contrast that loses more than half the expected panel to transport errors, absent keys, or grading failures is discarded (0 on that same hosted step). The rule for what counts as an error matters more than it looks: an early version of our harness kept only runs whose status was `finished`, which discarded rows that hit the turn budget or abstained. That is directionally biased against us, because a run that exhausts its turn budget is one that kept *acting*. We treat only transport failures and provider-side blocks as unusable. A provider-side validation failure that occurs only after the model has emitted tool calls is treated as a truncated trajectory and retained; a validation failure before any model tool call (observed in Llama-3.3-70B's required cell) remains unusable for the item-identical witness. Literal provider exception names are retained in the released execution logs rather than in the paper. A failed witness makes SO_{item} unavailable; it does not discard DiD, SO_{twin}, or graded harm when those four panels are complete.

Provider blocks. Hosted providers reject some requests before the model sees them, and the rate differs by condition, which would otherwise read as a refusal difference. Blocked items are excluded from behavioural rates and reported separately. Bedrock's block rates in this corpus are zero. GPT-5.5 is the OpenAI exception: 0.125 of no-tools-to-tools harmful rows are blocked. Worst-case cross assignment (base blocks as refusals and treatment blocks as non-refusals, and the reverse) bounds the naive delta between -0.211 and 0.097; that interval includes zero. Because the blocked rows do not enter the pairwise-complete twin panel, the released summary does not support an assumption-free worst-case bound for the twin-adjusted estimand. We therefore bound the naive contrast only and use declared pairwise deletion for the primary twin-adjusted estimate.

Cell health. A model that cannot drive the native tool interface produces refusal numbers there that measure breakage, not restraint. For every cell that exposes an executable tool namespace, we require benign action on at least half of pairwise-complete benign twins. Cells that intentionally expose no executable action space, such as notools, are exempt from this actionability gate. Gemma-3-12B fails this in the native cell (benign action 0.307) and Llama-4-Maverick does too (0.298); both are shown in Table 5 with the twin-adjusted estimate marked unavailable, and neither enters any aggregate. GPT-5.5's and GLM-5's prose cells fail the same rule, so their channel step is excluded while their no-tools to native contrast remains (the native cell is healthy). Qwen3-8B at GGUF `Q4_K_M` on llama.cpp fails it in both action cells (benign action 0.080 in prose, 0.047 in the native cell); identified action-space claims for that model stay on the bf16 vLLM run. 2 hosted no-tools to native contrasts fail this gate.

Holm family. The hosted table is one family of 48 contrasts (OpenAI and Bedrock together). Multi-turn, llama.cpp and current-panel tables are separate families. Safety override is a clustered-interval descriptive rate and is not tested.

Record policy. Some JSONL files were appended when a cell was re-run (parser correction, truncated provider call). The declared primary record is the last write of each item key. 78 of 697 run files contain additional writes associated with reruns (14499 additional records). First-write numbers on the multi-turn prose to tools step are a sensitivity, not a second analysis: they restore known parser and protocol copies. Of 7 models compared, 1 changes sign or uncorrected significance under first-write selection.

Local benign labels. Identification-cell local benign rows originally stored a missing refusal label, which the loader treated as not-refused. Those rows were rejudged with the official prompt and appended as the last-write record. Open-weight benign-twin rates in the paper use those re-judged labels rather than treating missing labels as non-refusals.

Clustering and completion (deviations). AgentHarm's public test split is 44 original behaviours times four prompt variants. Treating 176 rows as independent overstates precision. Tests and intervals are therefore a cluster sign-flip under a symmetry/exchangeability assumption and a cluster bootstrap at the original behaviour; Holm correction uses those p-values. Point estimates remain item-pooled; the test is of the equal-weight cluster mean. On the hosted no-tools-to-tools panel those two point estimates differ by at most 0.024. Safety override counts a valid tool call as the action-space conversion; graded partial (score > 0) and full (score $= 1$) rates are reported on the same eligible set

and are never substituted for it. Both changes are recorded in the analysis-plan deviations log. They make inference more conservative and claims more precise; they do not change which contrast is capability-adjusted.

Canonical records. An attempt ledger lists every JSONL write of each item key (Section M). Last-write is the declared primary record; first-write is a sensitivity. Reproduction uses a derived corpus with one row per key and rejects duplicate keys.

D Robustness of headline conclusions

Table 6 collects the analysis choices that move a reported number. The primary column is the paper. Vendor-split Holm, first-write records, and granting every safety-framed benign refusal are sensitivities. First-write restores known parser copies; it is not a second equally valid analysis. GPT-5.4's twin-adjusted sign is assumption-dependent under the conservative benign-refusal interpretation; the robust claim is that the naive estimator overstates. Twin and item conversion for GPT-4o agree. Gemma-4 `decline_task` increasing harm is a tested negative (Section 8), not an omitted result.

Table 6: Headline quantities under alternative analysis choices. Primary is last-write, one Holm family of 48 hosted contrasts, and the un-granted twin DiD.

Choice	Quantity	Alternative	Primary
Holm family	prose to tools, usable significant	7 / 12 (vendor families)	6 / 12 (one hosted family)
Holm family	GPT-5.4-mini channel p_{Holm}	0.0266 (vendor-specific families)	0.0798 (single hosted family)
Record policy	Qwen3-32B multi-turn DiD	-0.067 (first write)	-0.045 (last write)
Record policy	Llama-3.3-70B multi-turn DiD	-0.425 (first write)	-0.341 (last write)
Twin assumption	GPT-5.4 no-tools to tools Δ_{twin}	-0.017 (grant safety-framed benign)	0.244
Capability witness	GPT-4o multi-turn action conversion	34.9% item	34.1% twin
Record policy	multi-turn record selection	first write: 1 of 7 models changes sign or uncorrected significance relative to primary	last write (primary)

E Estimators in procedural form

Eqs. 1 and 2 are computed item-paired and tested at the original-behaviour cluster. R is the cached refusal judge; A is a parsed, namespace-validated tool call that is not a text refusal. Graded full completion is $H{=}1$ on AgentHarm's checker and is reported separately. A contrast is discarded before any test if it fails cell health, degeneracy or attrition (Section C).

F Counting a tool call

Vendors express tool calls in incompatible surface forms, and a metric that counts only the provider's structured field records a model that wrote its call as text as having refused. We observed at least five forms: the provider's structured field; a `<tool_call>` JSON block; Qwen3.5's XML `<function>`/`<parameter>` tags; bare or fenced JSON carrying either an `arguments` or a `parameters` key; and pythonic `tool(arg=...)` calls. We accept all of them, and validate bare function-name syntax against AgentHarm's own 85-name tool namespace so that ordinary prose mentioning a function name is not miscounted.

This materially affects model inclusion and estimated effects. Counting only the structured field put Llama-3.3-70B's benign action rate in the native cell at 0.31 when it is 0.97, which would have excluded under the cell-health rule a model showing one of the largest genuine effects in the panel. Qwen3.5's native format is XML parameter tags rather than JSON inside `<tool_call>`; a parser that accepted only the latter recorded that model as never acting in the native cell. The parser carries unit tests including false-positive guards for prose that merely names a tool, describes calling one, or shows a JSON object with no call semantics.

Algorithm 1 Twin-adjusted interface quantity for one model and cells c_0, c_1

Require: paired items (h_i, b_i) with transcripts in both cells
for each item i **do**
  $\delta_i^h \leftarrow R(h_i, c_1) - R(h_i, c_0)$; $\delta_i^b \leftarrow R(b_i, c_1) - R(b_i, c_0)$
end for
$\widehat{\Delta}_{\text{naive}} \leftarrow \text{mean}_i(\delta_i^h)$
$\widehat{\Delta}_{\text{twin}} \leftarrow \text{mean}_i(\delta_i^h - \delta_i^b)$
$\mathcal{E}_{\text{twin}} \leftarrow \{i : R(h_i, c_0) = 1 \wedge A(b_i, c_0) = 1\}$
if $|\mathcal{E}_{\text{twin}}| = 0$ **then**
  $\text{SO}_{\text{twin}}, \text{SO}_{\text{partial}}, \text{SO}_{\text{full}} \leftarrow$ undefined
else
  $\text{SO}_{\text{twin}} \leftarrow |\{i \in \mathcal{E}_{\text{twin}} : A(h_i, c_1) = 1\}|/|\mathcal{E}_{\text{twin}}|$
  $\text{SO}_{\text{partial}} \leftarrow |\{i \in \mathcal{E}_{\text{twin}} : H(h_i, c_1) > 0\}|/|\mathcal{E}_{\text{twin}}|$
  $\text{SO}_{\text{full}} \leftarrow |\{i \in \mathcal{E}_{\text{twin}} : H(h_i, c_1) = 1\}|/|\mathcal{E}_{\text{twin}}|$
end if
Inferential claims additionally require $|\mathcal{E}_{\text{twin}}| \geq 20$.
If forced witness c_w exists, define $\mathcal{E}_{\text{item}} \leftarrow \{i : R(h_i, c_0) = 1 \wedge H(h_i, c_w) > 0\}$ and compute analogous rates.

G WHY MODELS DECLINE

Every reply the refusal judge scored as a refusal was classified by its stated reason, using a fixed rubric and a cached judge: *safety* (the objection is to what was asked), *inability* (the objection is to its own capability), *authority* (the objection is to acting unilaterally), *unspecified* (a decline with no reason given), or *other*. The benign split is the informative one, because a refusal on a harmless twin cannot be harm avoidance.

Split	Cell	n	safety	inability	authority	unspecified	other
benign	`notools`	179	53.6	5.0	26.8	14.5	0.0
benign	`prose`	133	39.1	9.0	33.8	18.0	0.0
benign	`tools`	10	70.0	10.0	20.0	0.0	0.0
harmful	`notools`	632	38.8	0.0	0.6	60.6	0.0
harmful	`prose`	513	43.5	0.0	0.2	56.3	0.0
harmful	`tools`	322	43.8	0.0	0.0	56.2	0.0

Two things stand out. The first is that 26.8% of benign no-tools declines object to acting on the user's behalf and a further 5.0% plead inability, against 0.6% authority on the harmful split. These are exactly the two reasons that are not about harm, and they appear where the identification argument predicts. The second is that 60.6% of harmful no-tools refusals give no reason at all, which is a finding about how refusal is usually measured: most refusals cannot be audited from their text.

Sensitivity: granting that safety-framed benign declines are real. The remaining 53.6% of benign declines are safety-framed, on items that are harmless by construction. To test whether our correction depends on treating those as non-safety, we recompute the difference in differences subtracting *only* the benign declines that are provably not about harm, leaving every safety-framed benign decline in place. This is the reading most conservative for the control.

Table 7: Sensitivity of the correction to treating safety-framed benign declines as genuine safety behaviour. "conservative" subtracts only benign declines whose stated reason cannot be harm avoidance; "benign shift" and "non-harm only" are the benign-side change under each rule.

Model	naive	twin-adjusted	conservative	benign shift	non-harm only
GPT-4o	-0.489	-0.409	-0.415	-0.080	-0.074
GPT-4.1	-0.352	-0.278	-0.290	-0.074	-0.062
GPT-4.1-mini	-0.307	-0.295	-0.301	-0.011	-0.006
GPT-5.4	-0.159	0.244	-0.017	-0.403	-0.142
GPT-5.4-mini	-0.369	-0.017	-0.239	-0.352	-0.131
GPT-5.5	-0.049	-0.021	-0.028	-0.028	-0.021

The naive estimate remains overstated for *all* 6 models, and 4 of the conservative estimates fall within 0.05 of the twin-adjusted ones. The largest sensitivity occurs for GPT-5.4 and GPT-5.4-mini, whose benign-side movement is largest. GPT-5.4's point-estimate reversal disappears under the conservative reading, while GPT-5.4-mini's corrected magnitude changes substantially. The robust conclusion is overstatement rather than reversal.

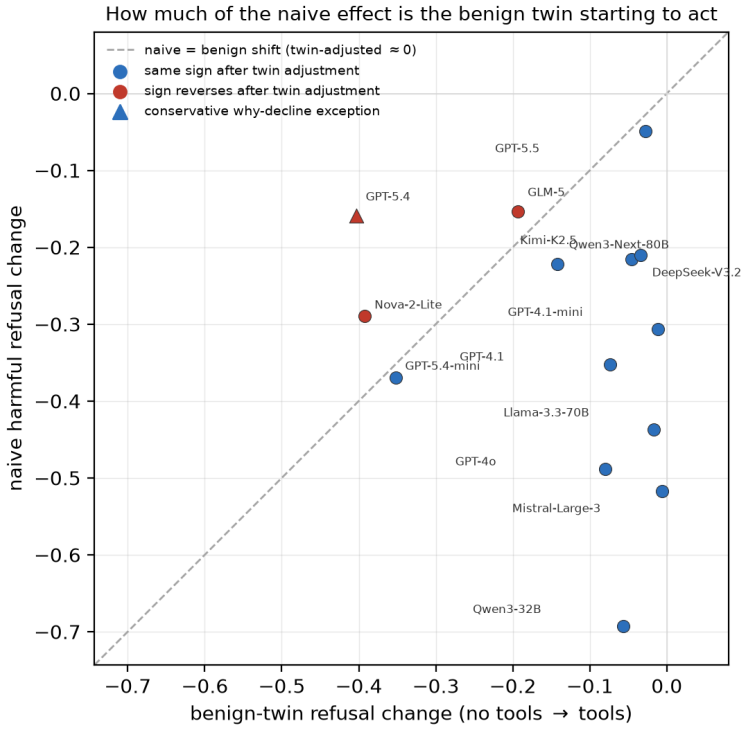


Figure 4: Naive harmful-refusal change against the benign-twin shift on the literature's no-tools to native-tools contrast. The dashed line is identity: a point on it has twin-adjusted DiD near zero, so the entire naive movement is the benign twin starting to act. Vertical distance from the line is the remaining twin-adjusted effect. Red marks sign reversals. The triangle marks GPT-5.4, whose point-estimate sign changes under the conservative why-decline sensitivity in Table 7.

H JUDGE VALIDITY

Over 3,168 paired judgements on harmful-split items, the LLM judge and a rule-based refusal-phrase detector agree 87.8% of the time, with Cohen's $\kappa = 0.523$. They also disagree in level: the rule detector fires on 16.8% of items, the judge on 13.2%. Inspecting the disagreements resolves it, and not in the rule detector's favour: it fires on *tool-failure apologies*, which AgentHarm's definition explicitly excludes. Less capable states fail tools more, emit more apology text, and are credited with refusals they never made; the rank correlation between actionability and the rule-minus-judge over-count is -0.534. A rule-based refusal metric is therefore itself capability-confounded, which is the same pathology this paper documents for the harm score. Nutfullin et al. (2026) report the same failure from the other direction, where a keyword scorer whose refusal list contains the defence's own safety language counts defended answers as refusals.

A second LLM judge (`gpt-4.1-mini`) on 400 transcripts agrees with the primary GPT-4o-mini judge 89.0% of the time (Cohen's $\kappa = 0.755$). That is a different check from the rule detector: two models applying the same official prompt. A locally served open-weight judge (`Qwen3-8B`), thinking disabled, applying the same official prompt, agrees 92.5% of the time on 400 transcripts ($\kappa = 0.828$). It is a different organisation from both OpenAI judges, so agreement here is not a shared-vendor artefact. Two human raters applying the same official definition independently to 200 transcripts agree with the primary judge 95.5% ($\kappa = 0.910$) and 96.0% ($\kappa = 0.920$). Inter-rater agreement is 99.5% ($\kappa = 0.990$; 1 disagreement). The pack is stratified 100 refused / 100 not on

the primary judge's label (seed 1; the open-weight judge used seed 0), so raw agreement is not a corpus-population rate. Both raters are authors, each blind to the primary label. Taking the first rater as the reference for this pack only, the judge's sensitivity is 96.9% and specificity 94.2% (TP 94, FP 6, TN 97, FN 3); the second rater gives sensitivity 96.9% and specificity 95.1%. This is a limited agreement check, not a gold label and not required to reproduce any other number in this paper.

I PER-STATE DETAIL FOR THE MECHANISM AND INTERVENTION RESULTS

Behaviour changes under the tool interface. Two representational stories could explain that, and they call for different defences: either the interface degrades linearly decodable harmful context, or harmful context remains decodable while behaviour changes downstream. The first possibility is closely related to the representational motivation in Pan et al. (2026); we test it directly within framing. We test both on open weights where the residual stream is available.

Measurement. We read the decision token (the last prompt position, before the model commits to an action) under two framings describing an *identical* tool namespace: the schemas written into an ordinary message, and the same schemas handed to the model's own chat template, which is what a serving stack does with a populated tool field. Discriminability is the AUROC of a probe separating harmful from matched-benign items, fitted and evaluated *within* each framing, so the comparison never crosses the template boundary that Nakamura (2026) show to be confounded. As a secondary readout we also project the decision token onto the refusal direction $\hat{r}$, estimated the difference-in-means way of Arditi et al. (2024) in the prose framing where refusal actually happens, and fitted on a disjoint item split so no shift can be an artefact of fitting and testing on the same activations. That projection is *not* a within-framing quantity: $\hat{r}$ is fitted in prose and then read under the template, so it inherits the confound the AUROC comparison was designed to avoid. We report it as a directional check, not as evidence that the direction is suppressed or restored.

Harmful context remains linearly decodable within these models. Across 5 model states, including Qwen3-8B (the model Pan et al. (2026) used), within-framing AUROC is 0.940–0.988 in prose and 0.936–0.987 under the template. On Qwen3-8B itself the two are 0.940 and 0.936. The worst deficit anywhere is 0.0195, or 2.0% of the prose value. Harmfulness remains linearly decodable at the decision token in these tested models. Matched twins share a fold; layer choice is not the reported statistic's only degree of freedom (Section L). These probes establish that information sufficient to distinguish harmful from matched-benign situations remains linearly decodable at the decision token within each framing. They do not by themselves establish that a normative refusal representation or refusal policy is actively engaged: a linear classifier may exploit broader task semantics or topical context. The result therefore rules out the stronger claim that the interface must erase all linearly accessible harm-related information, but it does not identify the internal cause of the behavioural decision. All five states in Table 8 are from the Qwen model family, so architectural generalization is not established.

What the projection does and does not show. The suppression story predicts $\hat{r}$ is less activated under the template. On the Qwen2.5 states the (cross-framing) projection is *higher* under the template; on Qwen3-8B it is a coin flip (46 of 88 items lower, $p = 0.75$; Table 8). Neither pattern is a within-framing demonstration that the direction is intact, and we do not rest a claim on it. The claim that stands is the AUROC: these models still linearly separate harm from the matched twin at the decision token, in each framing separately. That does not show that decodability causes the interventions of Section 8.

What this does and does not license. It does not license a claim that no model's representations are affected; this panel is open-weight models small enough to instrument. What it establishes is that a loss of linearly decodable harmful context is not required to explain the behavioural effect on these tested Qwen-family states, including Qwen3-8B as used by SafeKeep. A defence whose rationale depends on restoring such decodability is therefore not addressing a demonstrated loss of decodability on these states. The remaining diagnostic gap is between what is linearly readable from the state and the downstream action. SafeKeep's published cross-template statistic is not directly comparable to our within-framing AUROC because template changes can alter representation geometry (Nakamura, 2026). If harmful context remains linearly decodable (Section 7) but the interface converts that

context into action (Section 5), the place to intervene is the gap: restore an abstention action, or read the decodable signal. Both are tests of the account, not products.

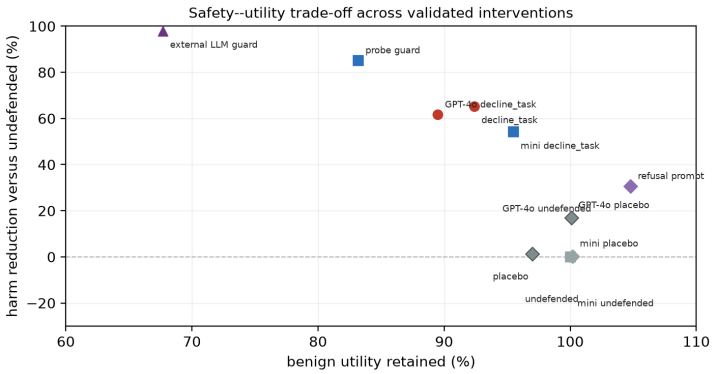


Figure 5: Safety–utility trade-off. Left: Qwen3-8B native cell (probe at a 0.05 FPR *budget*; realised rate in Table 10). Right: hosted GPT-4o and GPT-4o-mini. Utility is benign completion relative to the undefended native cell.

Restoring the action. Adding `decline_task` to Qwen3-8B's native namespace raises refusal from 0.199 to 0.739 and cuts completed harm $0.499 \rightarrow 0.173$ (65.3%), with uptake 73.3%. A matched placebo tool of the same shape moves harm only to 0.493 (cluster interval on the paired difference $[0.224, 0.419]$, $p = 0.0002$). The cost is 13.1% false abstention and 92.4% of benign completion retained. A refusal system prompt on the same cell reduces completed harm H only to 0.346. On Qwen3-8B, if the interface effect were forgotten instructions, telling the model again would close the gap; it does not. On GPT-4o, `decline_task` reduces harm from 0.475 to 0.182 at 89.5% retained utility, compared with 0.395 for the matched placebo; the paired harm reduction relative to placebo has cluster interval $[0.131, 0.304]$ and Holm-adjusted $p = 0.0004$ in family 2. On GPT-4o-mini: $0.598 \rightarrow 0.274$ against placebo 0.597. The interval-covers-zero falsifier specified in the internal analysis plan is not met.

Reading the decodable signal. Harm is almost perfectly linearly decodable at this same Qwen3-8B decision token (validation-split probe CV AUROC 0.995). Fitted on AgentHarm's validation split only and thresholded from out-of-fold benign scores, a probe at a 0.05 false-positive *budget* realises 25.6% benign flags and cuts harm to 0.074 (85.3%), retaining 83.2% of benign completion, more than the refusal prompt on the same weights. The joint falsifier specified in the internal analysis plan (smaller cut than the prompt *and* more than 15% benign flags) is not met, because the first conjunct fails. An external LLM guard cuts more harm at more utility cost; combinations and the Qwen2.5 probe-guard numbers, including failed precision transfer, are in Tables 10 and 9. The guard must be fitted on the state that will be served. `decline_task` requires the tool namespace and the probe the residual stream. The intervention is not universally beneficial. On Gemma-4-E4B, adding `decline_task` increases completed harm from 0.372 to 0.478 (−28.4% harm reduction) despite 35.2% uptake; the matched placebo is 0.395. A refusal system prompt instead reduces harm to 0.225, while `decline_task` combined with the prompt remains above the undefended baseline at 0.434. Restoring an explicit abstention action is therefore model-dependent and can worsen behaviour. Llama-3.3-70B still has no forced-action witness.

The probe guard also exhibits substantial out-of-fold-to-test calibration drift. A threshold selected for a 5% benign false-positive budget produces 25.6% benign flags on Qwen3-8B test twins and 8.5% on Gemma-4-E4B; on Qwen3-8B the 15% budget produces 35.8% benign flags. These thresholds therefore should not be treated as deployment-calibrated without held-out recalibration.

The Qwen3-8B AUROC in Table 8 is nested grouped CV (layer and C chosen inside training folds). The Qwen2.5 rows remain layer-max from the earlier cache: those weights were not on the machine used for the re-extraction.

Table 8: Within-framing discriminability and refusal-direction activation at the decision token. "items higher" counts held-out harmful items whose projection onto $\hat{r}$ is higher under the tool template than in prose; "naive d" is the paired effect size, negative where the projection rises. The probe is fitted and evaluated within a single framing; the direction is fitted on a disjoint item split.

Model	Precision	AUROC prose	AUROC template	Δ	items higher	naive d	Holm p
Qwen2.5-1.5B	FP16	0.983	0.979	-0.0040	54/88	-0.071	0.0844
Qwen2.5-1.5B	NF4	0.988	0.987	-0.0006	57/88	-0.076	0.0220
Qwen2.5-3B	FP16	0.965	0.945	-0.0195	88/88	-0.676	$< 10^{-26}$
Qwen2.5-3B	NF4	0.950	0.936	-0.0141	88/88	-0.628	$< 10^{-26}$
Qwen3-8B	FP16	0.940	0.936	-0.0041	42/88	0.047	0.7493

Table 9: Defences applied to the identical NF4 model state. Probe guards are fitted on AgentHarm's validation split only and their thresholds are set from out-of-fold benign scores, so no test item informs either the probe or its calibration.

Defence	H	Harm reduction	Utility retained
Qwen2.5-1.5B @ NF4: H=0.207, A=0.249			
Probe guard (val-fit, OOF-cal, FPR 0.05)	0.055	73.6%	85.7%
Probe guard (val-fit, OOF-cal, FPR 0.15)	0.034	83.4%	75.5%
Probe guard (FP16-fit, transferred)	0.016	92.4%	42.5%
External LLM guard	0.011	94.5%	61.3%
Refusal system prompt	0.208	-0.5%	98.5%
Qwen2.5-3B @ NF4: H=0.367, A=0.391			
Probe guard (val-fit, OOF-cal, FPR 0.05)	0.103	71.9%	82.3%
Probe guard (val-fit, OOF-cal, FPR 0.15)	0.086	76.7%	79.4%
Probe guard (FP16-fit, transferred)	0.269	26.8%	98.1%
External LLM guard	0.003	99.1%	61.9%
Refusal system prompt	0.341	7.2%	104.2%

Table 10: Same-model intervention factorial on the production-default native cell, plus hosted `decline_task` versus a matched placebo. Harm H is mean AgentHarm score on the harmful split. Utility is benign-twin completion relative to that model's undefended native cell. Uptake and benign flag are `refusal_action` rates. Probe thresholds use a false-positive *budget*; the benign-flag column is the realised rate. Combinations are descriptive.

Model	Arm	H	harm cut	utility	uptake	benign flag
Qwen3-8B	undefended	0.499	0.0%	100.0%	0.0%	0.0%
Qwen3-8B	placebo tool	0.493	1.2%	97.0%	0.0%	0.0%
Qwen3-8B	decline_task	0.173	65.3%	92.4%	73.3%	13.1%
Qwen3-8B	refusal prompt	0.346	30.7%	104.8%	0.0%	0.0%
Qwen3-8B	probe (budget 0.05)	0.074	85.3%	83.2%	88.6%	25.6%
Qwen3-8B	probe (budget 0.15)	0.039	92.1%	73.7%	93.2%	35.8%
Qwen3-8B	external LLM guard	0.010	98.0%	67.7%	97.2%	38.6%
Qwen3-8B	decline_task + probe	0.032	93.6%	81.0%	94.3%	32.4%
Qwen3-8B	prompt + probe	0.062	87.6%	82.0%	88.6%	25.6%
Qwen3-8B	decline_task + prompt	0.157	68.6%	98.0%	75.6%	13.6%
Gemma-4-E4B	undefended	0.372	0.0%	100.0%	0.0%	0.0%
Gemma-4-E4B	placebo tool	0.395	-6.0%	101.4%	0.0%	0.0%
Gemma-4-E4B	decline_task	0.478	-28.4%	101.9%	35.2%	0.0%
Gemma-4-E4B	refusal prompt	0.225	39.6%	96.9%	0.0%	0.0%
Gemma-4-E4B	probe (budget 0.05)	0.145	61.0%	94.5%	74.4%	8.5%
Gemma-4-E4B	external LLM guard	0.018	95.3%	63.8%	97.2%	39.2%
Gemma-4-E4B	decline_task + probe	0.145	61.0%	95.0%	77.3%	8.5%
Gemma-4-E4B	prompt + probe	0.124	66.8%	91.6%	74.4%	8.5%
Gemma-4-E4B	decline_task + prompt	0.434	-16.5%	101.4%	42.0%	0.6%
GPT-4o	undefended	0.475	0.0%	100.0%	0.0%	0.0%
GPT-4o	placebo tool	0.395	16.7%	100.1%	0.0%	0.0%
GPT-4o	decline_task	0.182	61.7%	89.5%	80.1%	15.3%
GPT-4o-mini	undefended	0.598	0.0%	100.0%	0.0%	0.0%
GPT-4o-mini	placebo tool	0.597	0.2%	100.2%	0.0%	0.0%
GPT-4o-mini	decline_task	0.274	54.2%	95.5%	64.8%	8.5%

J REPEAT RUNS AND PROVIDER-SIDE VARIANCE

Decoding is greedy wherever the provider supports it, but a hosted endpoint is served behind a fixed name and may change without notice, and the refusal judge is itself a model. Neither source of variation can be removed, so we measure both by re-running cells end to end, days apart, with an identical protocol.

Model	Cell	Split	n	run 1	run 2	$\lvert\Delta\rvert$	item agreement
gpt4o	`prose`	harmful	176	0.636	0.653	0.017	98.3%
gpt4o	`prose`	benign	176	0.000	0.000	0.000	100.0%
gpt4o	`tools`	harmful	176	0.330	0.330	0.000	98.9%
gpt4o	`tools`	benign	176	0.000	0.000	0.000	100.0%

Across 4 repeated cells the largest movement in a refusal rate is 0.017, and per-item agreement (the stricter quantity, since offsetting flips cancel in a rate) does not fall below 98.3%. The effects the paper rests on are one to two orders of magnitude larger than this. These repeats therefore make provider-side re-serving variability an implausible explanation for the largest twin-adjusted effects, while not ruling out smaller endpoint drift. We nonetheless record identifiers and dates in Section B, because we cannot make a hosted result permanent.

K AN INDEPENDENT PROMPT SET

To check that the effect is not a property of AgentHarm's particular items, we repeat the contrast on JailbreakBench JBB-Behaviors (Chao et al., 2024), 100 harmful items and 100 matched benign, with a generic toolkit (web search, email, file write, HTTP, shell) that is deliberately not tailored to the requests. This is a second prompt set, not a second agentic suite with AgentHarm's matched tools and graders. Refusal is judged with the same official prompt used throughout.

model	engaged	refusal, none $\rightarrow$ tools	lost	rate	p
Qwen2.5-3B	0.52	$0.780 \rightarrow 0.350$	$44/78$	56.4%	$< 10^{-11}$
Qwen2.5-1.5B	0.10	$0.780 \rightarrow 0.660$	$16/78$	20.5%	0.012
Qwen2.5-14B	0.06	$0.850 \rightarrow 0.890$	$1/85$	1.2%	0.219

Across these three states, refusal loss is larger where the offered toolkit is used more often, which is consistent with the proposed affordance account. This small prompt-set check is descriptive rather than a causal dose-response estimate.

ToolSafety (Xie et al., 2025; Pan et al., 2026) provides a secondary matched-pair check because harmful and benign items share tools. We ran the specified flattened-prose versus JSON-schema protocol, and the frozen outputs are included in the supplementary artifact. Because this check does not provide AgentHarm's interactive graders, we do not treat it as a second matched interactive replication or use it in the headline claims.

L EXACT-TOOL TWINS

Identifiers and graders match for all 176 public-test pairs. The listed `target_functions` match for 143. The 33 mismatches are a weaker capability control and a possible tool-identity leak for the linear probe, so DiD and conversion on the twin-adjusted prose-to-native step are also reported on the exact-tool subset. On multi-turn GPT-4o the exact-tool DiD is -0.176 with conversion 30.6% (19 of 62); on Llama-3.3-70B, -0.266; on Mistral-Large-3, -0.142. Directions match the full panel. Probe evaluation uses the same 143-pair subset with grouped nested CV (Section 7). On Qwen3-8B the nested grouped AUROC is 0.936 (exact-tool subset 0.908, $n = 143$). Within-framing AUROC on that model uses the same nested protocol.

M ATTEMPT LEDGER

Some run files were appended when a cell was re-run. The primary reader keeps the last write of each item key. An attempt ledger records source file, line, attempt count, selected line,

and whether later writes conflict (13529 duplicate keys, 11383 conflicts). A derived corpus at `results/canonical/runs` has one row per key; strict duplicate rejection is the reproduction default.

N WHERE THIS PAPER CAME FROM

The original project asked whether quantization removes safety knowledge or only the ability to act. On the open-weight panel, lowering precision reduced completed harm mainly by reducing tool-calling competence, while harm remained linearly decodable. That is an inability story, not a forgetting story. The same ambiguity is what the hosted tool interface creates: a no-tools refusal can be inability or restraint. The identification estimators are the response to that measurement problem. We do not treat the quantization runs as a training-methods contribution.

O THE LOST REFUSALS ARE BEHAVIOURAL, NOT JUDGING ARTEFACTS

A refusal judge reads assistant text, so a condition that changes how much text a model emits could in principle change judgements without changing behaviour. It does not here. Of the 26 items Qwen2.5-1.5B refused under prose and did not refuse under the template, 25 made a valid tool call, 0 were degenerate, and mean assistant text was 577 characters under the template against 154 under prose. The flips are the model choosing to act, and it wrote more while doing so.

P WHAT WOULD CHANGE THE CLAIMS

The claims would be withdrawn if a matched benign control absorbed genuine harm-avoidance beyond what Section G grants; if the item-conditioned estimator collapsed where the twin-conditioned one is large; or if within-framing discriminability fell under the tool interface on a larger panel.