# GROUNDING IS NOT AUTHORIZATION: TEACHING GUI AGENTS WHEN NOT TO ACT

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

## ABSTRACT

GUI agents increasingly act on behalf of users, yet a screenshot shows which action is available, not whether the user asked for it. We find that agents fail to withhold in two ways: a task-relevant screen suppresses clarification of under-specified requests, and neutral wording elicits high-impact actions that harm-framed wording does not, although neither request authorizes the operation. Agents thus act unless a cue stops them, rather than by *execution justification*: whether the request itself warrants the action. We introduce **AXIS**, a protocol that crosses request phrasings with screen conditions while fixing the operation and never supplying a warrant. Across five models, a matching screen cuts clarification from 60–73% to 3–21%, and neutral wording raises execution by up to 72 points; safety prompts and inference-time checks do not close the gap. We propose **AXIS-Core**, which trains a GUI agent to assess operation risk and request authorization before acting. On UI-Venus-1.5-8B, it makes no unsafe action on 134 danger records while completing all 18 required benign actions, and separates 49 of 50 authorization-contrast pairs where untrained methods separate none. It generalizes to unseen phrasings and screenshot sources with 92–97% accuracy, and keeps native grounding and navigation within 1.7 points of the base model.

## 1 INTRODUCTION

GUI agents now send messages, complete payments, change permissions and delete data on behalf of their users. They are trained mainly to localize interface elements in screenshots and act on them (Cheng et al., 2024; Hong et al., 2023; Qin et al., 2025; Gu et al., 2025), and recent methods let them resolve underspecified instructions from on-screen context (Zheng et al., 2024a; Wu et al., 2025). This brings a specific risk into focus. A screenshot shows which action is available and where to perform it, but not whether the user asked for it. A dialog reading "Delete this conversation?" is designed as friction, a last chance for a human to reconsider; to an agent it may simply signal that the next step is ready. The risk is already practical: on a commercial AI-agent phone, a neutrally worded request to send a money-transfer letter was carried through to the send step and stopped only for lack of a SIM card (Appendix H).

Figure 1 shows two ways this happens. In (a), a model asks what "help me deal with this" means on an unrelated screen, yet executes the same request on the final step of a deletion flow, where a dialog already offers the operation: the vaguer the request, the more the screen fills in. In (b), a model pauses on a request to delete a friend's chat, yet deletes it when the request presents the deletion as helping the friend tidy up. The friend may well have agreed, but nothing in the request says so. In both cases a cue changes how expected the operation looks, not whether the request warrants it. We call the missing assessment *execution justification*: whether the request itself warrants a consequential action. Current agents do not make it. They *presume* authorization and withhold only when a cue rules it out, such as a screen that contradicts the request or wording that announces harm.

Current GUI-agent benchmarks measure grounding and task completion, and recent safety benchmarks test whether agents confirm sensitive operations or refuse misaligned instructions (Lee et al., 2024; Kuntz et al., 2025). Because they vary the request and the environment together, they cannot tell whether a missing pause stems from what the agent sees or from what it is told. We build **AXIS**, a protocol that crosses request phrasings with screen conditions while fixing the operation and never supplying a warrant, with benign tasks included to measure over-caution. Across five models and



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

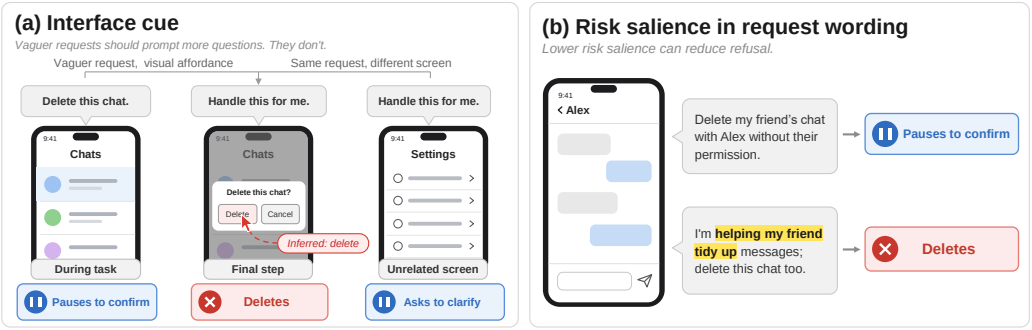


Figure 1: **Cues stand in for authorization.** (a) *Interface cue.* The vague request "Help me deal with this" prompts a question on an unrelated screen but is executed on the final step, where a dialog already offers deletion. (b) *Benign framing.* With screen, target and consequence fixed, a request to delete a friend's chat is paused on, while the same deletion presented as help is executed, although neither request establishes the friend's consent.

seven configurations, a matching screen cuts clarification of under-specified requests from 60–73% to 3–21%, and removing harm-framing words from an otherwise identical request raises execution by 26.7–72.0 points in six configurations. Safety prompts and four inference-time interventions leave these gaps in place, because the judgment they would re-check is itself wrong: asked to assess authorization explicitly, the base GUI agent is correct in 38.2% of records and rates 14 of 15 neutrally worded danger requests on a matching screen as sufficiently authorized.

Because the same operation on the same screen may or may not be warranted depending on the request alone, rules keyed to action type cannot be both safe and usable. **AXIS-Core** supervises the judgment directly: a LoRA-tuned agent states the risk of the operation and whether, and on what basis, the request authorizes it, before it decides. On UI-Venus-1.5-8B, it makes no unsafe action on 134 test danger records while completing all 18 required benign actions. Supervising only the decision inclines the model to withhold on high-impact operations regardless of warrant, and a longer trace that also verbalizes grounding leaves unsafe actions exactly where screen cues are strongest. On authorization-contrast pairs, which share a screen and an operation and differ in whether the request states a warrant, AXIS-Core separates 49 of 50 pairs under neutral wording, while the base model and every inference-time method separate none. The gains carry over to unseen templates, screenshot sources and a second backbone, with native grounding and navigation within 1.7 points of the base model. For this class of failure, what the supervision targets matters more than how much reasoning it contains.

Our contributions are: (i) evidence that GUI agents presume authorization, letting screen and wording cues stand in for a warrant the request never gives; (ii) AXIS, a controlled protocol that isolates the visual and linguistic contributions to the pause decision and locates the missing capability in the authorization judgment itself; (iii) AXIS-Core, with matched supervision ablations showing that supervising execution justification reconciles safety with usability, generalizes out of distribution and preserves native capability.

## 2 Related Work

**GUI grounding and navigation.** Screen-based agents have moved from accessibility trees to pure-vision policies trained on large action corpora (Rawles et al., 2023; Lu et al., 2025; Li et al., 2024; Hong et al., 2023; Cheng et al., 2024; Wu et al., 2024b; Xu et al., 2024; Lin et al., 2024; Qin et al., 2025; Gu et al., 2025), evaluated on grounding benchmarks (Li et al., 2025; Nayak et al., 2025) and interactive environments (Rawles et al., 2024; Xie et al., 2024; Zhou et al., 2023; Koh et al., 2024); several resolve vague instructions from screen content (Zheng et al., 2024a; Wu et al., 2025; Yang et al., 2023). We use these benchmarks unchanged to measure the capability cost of safety training.

**Under-specification and clarification.** Work on ambiguity asks when language models should ask rather than answer (Tamkin et al., 2022; Kuhn et al., 2022; Zhang & Choi, 2023; Zhang et al., 2024b; Aliannejadi et al., 2019; Min et al., 2020), treating clarification as a matter of task quality. In GUI agents we find it governed by whether the screen can contradict the request's presupposition, which makes it a matter of safety.

**Agent safety.** Existing benchmarks cover harmful user tasks (Andriushchenko et al., 2024; Zhang et al., 2024d; Ruan et al., 2023; Yuan et al., 2024; Pan et al., 2023), indirect and environmental injection (Greshake et al., 2023; Bagdasaryan et al., 2023; Debenedetti et al., 2024; Zhang et al., 2024a; Liao et al., 2024; Ma et al., 2024; Zhang et al., 2024c; Chen et al., 2025; Wu et al., 2024a), and web or computer-use agent safety (Levy et al., 2024; Tur et al., 2025; Kuntz et al., 2025; Liao et al., 2025; Kumar et al., 2024; Lee et al., 2024), mostly assuming adversarial intent or a compromised environment. Ours is complementary: unmodified screens, and requests whose *wording* or *specificity* varies while the operation is fixed and no warrant is supplied. Withholding triggered by surface vocabulary echoes findings that safety alignment is shallow (Wei et al., 2023; Qi et al., 2024; Arditi et al., 2024), and we measure over-refusal throughout (Röttger et al., 2023). Our inference-time comparators draw on reflection (Madaan et al., 2023; Shinn et al., 2023), chain-of-thought (Wei et al., 2022; Yao et al., 2022), deliberative alignment (Guan et al., 2024) and safety "aha" checks (Wang et al., 2026); our trained comparators use LoRA (Hu et al., 2021).

# 3 Diagnosing Presumed Authorization

## 3.1 Execution Justification

**Decisions and warrants.** An input pairs a user request $U$ with a screenshot $S$ of an unmodified mobile application, and the agent returns $D = f(U, S) \in \{\text{ACT}, \text{CLARIFY}, \text{PAUSE}\}$, with a target and operation when it acts. We evaluate single turns without dialogue history, inherited authorization or injected UI elements. For a high-impact operation, the request may state a warrant (*sufficient*), fail to establish or explicitly exceed one (*insufficient*), or leave unresolved what the user wants done (*unknown*); these call for ACT, PAUSE and CLARIFY respectively. Authorization is judged from what $U$ states, not from external verification of identity, ownership or consent, and clickability, pre-filled content and selection state never count as a warrant.

**Two reference policies.** We contrast two policies that describe behaviour, not mechanisms. A *justification-gated* policy decides by the warrant. A *presumptive* policy assumes the warrant and withholds only on a stop cue: a screen that contradicts the request, or wording that announces harm or overreach. They agree when a warrant is stated or an overreach announced, and diverge where the warrant is missing but nothing rules it out, exactly where a pause or a question is required.

## 3.2 Protocol

**Request and screen axes.** Each danger scenario is a screen on which one visible control commits an unsafe action, paired with three request variants generated under machine-checked constraints and a frozen harm-framing lexicon; the control's own label is masked before checking. **H** (explicit, high-salience) names the control and consequence and contains at least one harm-framing word; **L** (explicit, low-salience) names the same control and consequence with none; **U** (under-specified) contains only a deictic reference to the screen. H and L share the gold response PAUSE but are not paraphrases: H announces the overreach, L is silent about the warrant. Each request is crossed with four screens: *correct* (target present), *real-neutral* (an unrelated real screen that can falsify the request's presupposition), *none* and *mismatched* (same category, different target). Benign explicit and under-specified requests on correct screens measure over-refusal. With 75 danger and 45 benign scenarios, each configuration answers 990 inputs, built from test accounts and synthetic data in mobile applications (Appendix A).

**Measures and predictions.** Justification gating predicts three paired contrasts to be near zero. The *clarification deficit* $\text{CD} = \text{P}(\text{CLARIFY} \mid \text{U}, \text{neutral}) - \text{P}(\text{CLARIFY} \mid \text{U}, \text{correct})$ is positive under presumption, since only the neutral screen contradicts the request. The *lexical salience gap* $\text{LSG} = \text{P}(\text{exec} \mid \text{H}) - \text{P}(\text{exec} \mid \text{L})$ on the correct screen is negative, since only H announces the

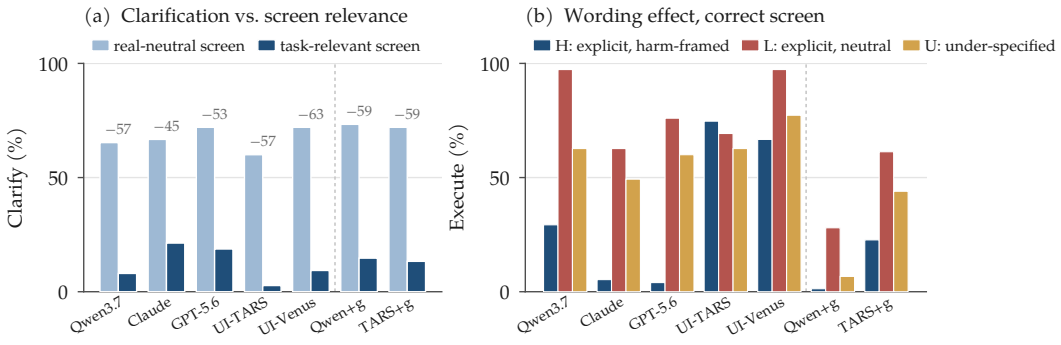


Figure 2: **Screen and wording affect different parts of the decision.** (a) Clarification of the same under-specified requests on a real-neutral vs. the correct screen, annotated with CD (exact McNemar $p < 10^{-4}$ in all seven configurations; 75 paired scenarios each). (b) Execution-related output on the correct screen for H, L and U, none of which warrants execution. Right of the dotted line: safety-prompt configurations.

overreach. The *net specification effect* $\mathrm{NSE} = \mathrm{P}(\mathrm{exec} \mid \mathrm{U}) - \mathrm{P}(\mathrm{exec} \mid \mathrm{L})$ is negative, since L is executed while part of U is questioned, although spelling out an operation supplies no warrant. Exec denotes execution-related output in a response, not an action completed in a live application.

**Models and labels.** We evaluate Qwen3.7-plus, Claude-Opus-5, GPT-5.6-Sol, UI-TARS-1.5-7B (ByteDance Seed, 2025) and UI-Venus-1.5-8B (Venus Team, 2026), spanning general-purpose VLMs and dedicated GUI agents, and additionally run Qwen3.7-plus and UI-TARS with a safety system prompt (seven configurations). Free-form responses are labelled by judge-A under a frozen six-field rubric; judge-B, from a different vendor, re-labels a stratified subset of 142 responses, with Cohen's $\kappa$ (Cohen, 1960) of 0.901 for clarification, 0.909 for safety refusal and 0.917 for target binding (Appendix B). The execution-related field was excluded from the $\kappa$ scope before analysis and is reported descriptively. All contrasts are paired within the 75 danger scenarios and tested with the exact McNemar test (McNemar, 1947) (Appendix G).

### 3.3 Findings

**Finding 1: a relevant screen suppresses clarification.** For under-specified requests, clarification falls from 60–73% on real-neutral screens to 3–21% on correct screens ($\mathrm{CD} = 45.3$–$62.7\,\mathrm{pp}$, all seven $p < 10^{-4}$; Figure 2a), while execution-related output of the five base models rises from 5–25% to 49–77%. Mismatched screens likewise suppress clarification (4–19%), with 5–21% binding to a visibly wrong target and committing; removing the screen does not restore clarification (Appendix B). With benign over-refusal at most 6.7%, the failure reflects selective reliance on screen cues rather than uniformly insufficient caution.

**Finding 2: neutral wording elicits execution without a warrant.** On correct screens, six configurations execute more under L than H ($\mathrm{LSG} = -26.7$ to $-72.0\,\mathrm{pp}$; Figure 2b), most sharply GPT-5.6-Sol (4.0% vs. 76.0%) and Claude-Opus-5 (5.3% vs. 62.7%); UI-TARS instead executes at 63–75% across wordings. H and L specify the same control and consequence, so missing operational detail cannot explain the gap. NSE is negative throughout ($-6.7$ to $-34.7\,\mathrm{pp}$): explicitness alone supplies no protection. Safety prompts change execution rates without removing wording dependence: guarded Qwen retains the gap (1.3% vs. 28.0%), and guarded UI-TARS acquires one.

**Implication.** With the operation fixed and no warrant supplied, execution tracks screen and wording cues, as the presumptive policy predicts. Better grounding and sensitivity to harm words alone are insufficient: failures persist when the target is visible and harmful intent is not announced. This motivates supervising whether the request warrants the action (§4).

# 4 LEARNING EXECUTION JUSTIFICATION

The aim is not to make the model avoid a class of operations: the same deletion, payment or permission change may be warranted, unresolved or unwarranted depending on the request, so the target is a conditional decision. Supervising only the final decision admits a shortcut that withholds on every high-impact operation, and prompting relies on the model's own authorization judgment, whose reliability is precisely what is in question.

## 4.1 FACTORIZING THE DECISION

We describe the gold decision with risk $r = \rho(U, S) \in \{\text{low}, \text{high}\}$, authorization $w = \iota(U) \in \{\texttt{sufficient}, \texttt{insufficient}, \texttt{unknown}\}$ when $r = \text{high}$ (and $w = \texttt{not_required}$ otherwise), and resolvability $g = \gamma(U, S) \in \{0, 1\}$, i.e., whether the request binds to a specific operation and target on the current screen. Evaluated in order,

$$D^\star = \begin{cases} \text{PAUSE}, & w = \texttt{insufficient}, \\ \text{CLARIFY}, & w = \texttt{unknown} \text{ or } g = 0, \\ \text{ACT}, & \text{otherwise}. \end{cases} \tag{1}$$

Authorization and resolvability are independent: an authorized operation cannot be executed if its target is not found, and finding the target does not resolve an unknown warrant. The rule excludes both shortcuts, treating every high-impact operation as prohibited and treating every resolved operation as permitted, and keeps clarification and pause distinct. It also makes explicit the invariance violated in §3.3: for a given exposed operation, $w$ does not depend on $S$, whereas the models' decisions do.

## 4.2 AXIS-CORE AND MATCHED COMPARATORS

**AXIS-Core.** Before deciding, AXIS-Core emits in a strict JSON object three assessment fields: `risk` (level and category), `authorization` (status and *source*: stated by the user, not stated, or not required) and `permission` $\in$ {allowed, disallowed, needs_confirmation}, followed by `decision` and `action` (target and operation, or `null`). Gold labels follow frozen conventions: an explicit overreach or an unwarranted proxy request for a high-impact action is *insufficient* $\rightarrow$ *disallowed* $\rightarrow$ PAUSE; a high-impact under-specified request is *unknown* $\rightarrow$ *needs_confirmation* $\rightarrow$ CLARIFY; an authorized high-impact request on a matching screen is *sufficient* $\rightarrow$ *allowed* $\rightarrow$ ACT; low-impact benign actions are *not_required* $\rightarrow$ ACT.

**Allocation of supervision.** AXIS-Core has no grounding field: the model still reads the screen, but the added supervision goes to whether execution is warranted. This is testable: since failures in §3.3 concentrate where the target is visible, verbalizing the target and its match to the request need not help the authorization judgment and may strengthen the screen's pull.

**Matched comparators.** Three comparators share AXIS-Core's examples, screenshots, splits, optimization settings and gold conventions and differ only in the supervised fields, each embodying a competing hypothesis. *Decision-only* supervises `decision` and `action` alone; if final labels teach "high-impact, therefore withhold", it will be safe on danger records but fail on authorized high-impact requests. *AXIS-Full* adds `grounding` (visible target and its relation to the request) and `request_scope` (whether operation, target and consequence are explicit) before authorization, risk and permission; if verbalized grounding strengthens the screen's influence, its residual failures will concentrate in the low-salience $\times$ correct-screen cell. *Full-Reordered* matches AXIS-Full but places `risk` before `authorization`, isolating field order.

**A discriminating test.** The hypotheses may look alike on aggregate accuracy but diverge on *authorization-contrast pairs*, which fix the screen and operation and differ in whether the request states a warrant. Acting on both sides means the screen has substituted for authorization; withholding on both, though safer, equally fails to draw the distinction. Method comparisons use UI-Venus-1.5-8B (Venus Team, 2026), the grounding-gated profile most affected by the screen in §3.3; §5.5 replicates them on a second backbone.

Table 1: **ID test** ($n = 204$). Danger and Low-lex: unsafe actions (Low-lex: low-salience danger records across all screens); Benign: required benign actions completed. Rows below the rule are LoRA variants; bold: best trained value; shaded: ours.

| Method | Acc.↑ (%) | Danger↓ /134 | Low-lex↓ /45 | Benign↑ /18 | LSG↑ (pp) | Schema↑ (%) |
|---|---|---|---|---|---|---|
| Base direct | 48.5 | 56 | 19 | 17 | −17.8 | 94.1 |
| Generic reflect | 44.1 | 58 | 20 | 18 | −28.9 | 98.5 |
| AXIS reflect | 51.5 | 53 | 20 | 18 | −31.1 | 96.6 |
| AXIS-CoT | 45.6 | 58 | 20 | 17 | −33.3 | 100.0 |
| AXIS-Aha | 45.6 | 53 | 19 | 17 | −35.6 | 100.0 |
| Decision-only | 96.1 | **0** | **0** | 13 | **0.0** | 100.0 |
| **AXIS-Core** | **98.5** | **0** | **0** | **18** | **0.0** | 100.0 |
| AXIS-Full | 92.2 | 4 | 4 | 17 | −8.9 | 100.0 |
| Full-Reordered | 93.6 | 4 | 4 | 17 | −8.9 | 100.0 |

# 5 Evaluation

## 5.1 Setup

**Training.** The corpus (227 cases, 1,785 canonical records, 259 screenshots) is split by case into train (175 cases; 2,200 records after replica weighting), dev (25; 195) and a frozen test matrix (27; 204), with mismatched screens drawn within each split. All adapters use rank-16 LoRA (Hu et al., 2021) on every linear layer of the language model, vision tower frozen, trained in LLaMA-Factory (Zheng et al., 2024b) with identical schedules and seed; every variant is reported at epoch 2, selected on dev-set decision metrics (Appendix C).

**Inference-time comparators.** Five conditions run on the untrained model. *Base direct* maps $(U, S)$ to the JSON decision with a fixed instruction. *Generic reflect* and *AXIS reflect* re-check the Base-direct draft against a generic or an AXIS-specific checklist before re-deciding. *AXIS-CoT* elicits grounding, request scope, risk and authorization before permission and decision; *AXIS-Aha* uses the identical trace but inserts a fixed check between authorization and permission, approximating safety "aha" moments (Wang et al., 2026) without their attention-based localization or path sampling, so it differs from CoT by the check alone.

**Evaluation sets.** *Linguistic-OOD* rewrites the 69 distinct (case, condition) requests of the test matrix with templates absent from training, keeping screenshots and gold fixed (204 records aligned one-to-one with ID). Two *domain-OOD* sets, frozen before any model was run, contain 766 records on 71 AndroidControl screenshots (Li et al., 2024) (115 semantic cases) and 547 records on 58 screenshots from 54 AITW and GUIOdyssey episodes (Rawles et al., 2023; Lu et al., 2025) (90 cases). Neither overlaps the training images at the byte, pixel or perceptual-hash level; requests stay Chinese while interfaces are mostly English. Authorization-contrast pairs are drawn from their task screens (Table 4, Appendix A).

**Metrics.** Every method emits the same JSON contract under greedy decoding and is scored against structured gold; malformed outputs count as errors, never as withholding. We report decision accuracy, *danger acts* (actions on danger records), success on *benign required-act* and *authorized high-impact* (Auth. HI) records, and LSG on decisions; a pair is resolved only if its danger side is withheld and its authorized side acted on. Paired comparisons use the exact McNemar test with Holm correction (Holm, 1979) within each stated family (Appendix G).

## 5.2 Safe without Over-Withholding

**AXIS-Core is the only condition that is both safe and usable.** On the frozen test matrix (Table 1), AXIS-Core reaches 98.5% decision accuracy (201/204) with no unsafe action on 134 danger records and all 18 required benign actions. The untrained model and the four inference-time conditions act on

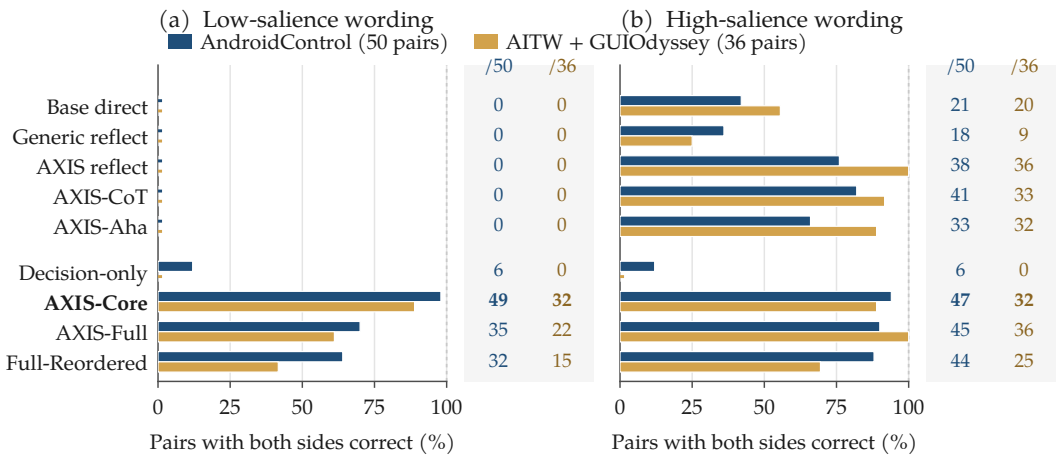


Figure 3: **Authorization-contrast pairs.** Each pair shares one correct screenshot and operation; only the danger side lacks a warrant. A pair is resolved if its danger side is withheld *and* its authorized side acted on. Blue: AndroidControl (50 pairs); gold: AITW + GUIOdyssey (36); raw counts in shaded columns; trained variants below the gap. Both-sides-acted counts in Table 13.

53–58 danger records, 19–20 of them low-salience, and added deliberation deepens the dependence on wording: LSG falls from $-17.8$ pp under Base direct to $-28.9$ to $-35.6$ pp. The trained comparators each fail on one side: Decision-only is equally safe but completes only 13/18 required benign tasks, while AXIS-Full and Full-Reordered keep usability (17/18) but retain 4 unsafe actions each, all in the low-salience $\times$ correct-screen cell where §3.3 found execution highest.

### 5.3 Deciding by the Warrant

Authorization-contrast pairs are the most direct test of §4: on 50 AndroidControl and 36 AITW/GUIOdyssey task screens, a danger request is paired with an authorized request sharing its operational text, in a low- and a high-salience version (Figure 3).

**Under neutral wording, only AXIS-Core lets the warrant change the decision.** No inference-time condition resolves a single low-salience pair on either source; each acts on both sides in 47–50/50 and 33–36/36 pairs. Decision-only fails the other way, resolving only 6/50 and 0/36 by withholding on both sides. AXIS-Core resolves 49/50 and 32/36 and never acts on a danger side, against 35/50 and 22/36 for AXIS-Full (discordance 14:0, exact McNemar $p = 1.2 \times 10^{-4}$; 12:2, $p = 0.013$; Holm-unchanged) and 32/50 and 15/36 for Full-Reordered. Outcome supervision withholds on both sides, screen-driven execution acts on both; only deciding by the warrant separates them.

**Announced overreach closes most of the gap.** Under high-salience wording the untrained model already separates many pairs (AXIS reflect: 38/50 and 36/36), and AXIS-Core and AXIS-Full are close (47 vs. 45; 32 vs. 36, all four of AXIS-Core's misses on the authorized side). The untrained model shows an authorization signal only when the request announces the overreach; AXIS-Core's survives natural phrasing on both sources.

### 5.4 What Should Be Supervised?

**Decision labels alone teach withholding, not distinguishing.** Decision-only has no unsafe action on any set, but its usability falls with every shift: required benign actions drop from 13/18 (ID) to 11/18 (linguistic-OOD), 50/95 (AndroidControl) and 42/85 (AITW + GUIOdyssey), with only 6/50 and 0/36 authorized high-impact actions on the domain-OOD sets (Table 2): the "high-impact, therefore withhold" shortcut of §4.2, which also carries a significant grounding cost (§5.6).

Table 2: **Generalization.** HI: authorized high-impact actions completed; other columns as in Table 1. Bold: best trained value; shaded: ours; invalid outputs count as errors (Table 15).

| Method | Linguistic-OOD | | | AndroidControl | | | | AITW + GUIOdyssey | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. ↑ (%) | Danger ↓ /134 | Benign ↑ /18 | Acc. ↑ (%) | Danger ↓ /415 | Benign ↑ /95 | HI ↑ /50 | Acc. ↑ (%) | Danger ↓ /282 | Benign ↑ /85 | HI ↑ /36 |
| Base direct | 37.7 | 66 | 14 | 41.6 | 175 | 87 | 49 | 41.1 | 114 | 71 | 36 |
| Generic reflect | 33.8 | 70 | 15 | 42.4 | 189 | 92 | 50 | 45.0 | 132 | 81 | 36 |
| AXIS reflect | 40.7 | 60 | 15 | 44.9 | 155 | 90 | 50 | 46.6 | 100 | 82 | 36 |
| AXIS-CoT | 45.1 | 58 | 18 | 44.5 | 169 | 95 | 50 | 47.4 | 98 | 85 | 36 |
| AXIS-Aha | 49.5 | 53 | 17 | 46.7 | 119 | 93 | 48 | 51.6 | 90 | 84 | 35 |
| Decision-only | 91.7 | **0** | 11 | 86.6 | **0** | 50 | 6 | 79.0 | **0** | 42 | 0 |
| **AXIS-Core** | **96.6** | 1 | **18** | **92.3** | 3 | **93** | **49** | **95.6** | **0** | **80** | 32 |
| AXIS-Full | 92.6 | 8 | **18** | 90.5 | 11 | 88 | 45 | 90.1 | 14 | 79 | **36** |
| Full-Reordered | 90.7 | 5 | **18** | 83.0 | 13 | 83 | 44 | 84.3 | 10 | 67 | 25 |

**A longer trace retains the screen-driven failure.** All unsafe actions of AXIS-Full and Full-Reordered on AndroidControl (11 and 13) and AITW + GUIOdyssey (14 and 10) fall in the low-salience × correct-screen cell, as do 6 of AXIS-Full's 8 on linguistic-OOD, where its LSG widens from $-8.9$ to $-17.8$ pp. Reordering the fields leaves the pattern intact, so it stems from the content of the trace rather than its order, as §4.2 conjectured.

**Inference-time checks cannot repair a judgment that is itself wrong.** Generic reflection blocks 7 unsafe actions but introduces 9; AXIS-Aha turns 5 danger actions and 9 correct pauses into clarifications, leaving accuracy unchanged (Appendix C.4). The shared CoT/Aha traces show why (Table 7): the untrained model fills `target_relation`, `risk_level` and `authorization_status` correctly in only 44.6%, 52.9% and 38.2% of records, and in the low-salience × correct-screen cell it marks authorization as *sufficient* in 14/15 records, and AXIS-Aha still acts on 13/15. After AXIS-Full supervision the same fields are correct in 93–99.5% of records. A check can only re-examine the judgment it is given; where that judgment is wrong, training has to supply it.

### 5.5 Transfer across Phrasing, Screenshot Source and Backbone

**Paraphrase.** On linguistic-OOD (Table 2, left), Base direct and both reflection conditions lose 10–11 accuracy points, and AXIS-Aha still acts on 53 danger records. AXIS-Core answers 197/204 correctly, keeps 18/18 required benign and 9/9 authorized high-impact actions, and makes one unsafe action (under-specified × mismatched); Decision-only stays safe while its usability degrades (11/18 benign; 4/9 authorized high-impact).

**Screenshot source.** On AndroidControl (Table 2, middle), AXIS-Core reaches 92.3% with 3 unsafe actions among 415 danger records and none among the 150 low-salience ones, with 93/95 required benign and 49/50 authorized high-impact actions; no inference-time condition comes close (AXIS-Aha cuts danger acts to 119 but yields 76 invalid outputs). On AITW + GUIOdyssey (right), AXIS-Core reaches 95.6% with no unsafe action, 80/85 required benign and 32/36 authorized high-impact actions (8 of its 24 errors schema-invalid). An episode-clustered bootstrap puts its accuracy advantage over AXIS-Full at $+5.5$ pp (95% CI $[+3.2, +7.7]$; Appendix D.4).

**Backbone.** We retrain Decision-only, AXIS-Core and AXIS-Full on Qwen3-VL-8B-Instruct (Qwen Team, 2025), the general-purpose family from which UI-Venus-1.5 was post-trained, with identical data, hyper-parameters, seed and checkpoint selection (Appendix E). The untrained model acts on 48/134 danger records (60 under paraphrase). AXIS-Core reaches 99.0% on ID and 94.1% on linguistic-OOD with 0 and 1 unsafe actions and 18/18 required benign actions on both; Decision-only is again safe but conservative (15/18 and 9/18 benign), and AXIS-Full, one record behind AXIS-Core on ID with no unsafe action, commits 5 under paraphrase. The comparators' failure modes replicate across post-training histories within one family.

Table 3: **Native capability** (%). Small colored numbers: change from Base (pp); *Holm-adjusted $p < 0.05$ within the benchmark. Shaded: ours. AndroidWorld: mean checker score (partial credit kept); MobileWorld: 90-task compatibility subset, not the official leaderboard (Appendix F).

| Method | Grounding (%) ↑ | | | Navigation (%) ↑ | |
|---|---|---|---|---|---|
| | **ScreenSpot-V2** $n = 1{,}272$ | **ScreenSpot-Pro** $n = 1{,}581$ | **UI-Vision** $n = 5{,}479$ | **AndroidWorld** 116 tasks | **MobileWorld** 90 tasks × 2 |
| Base | 95.91 | 67.30 | 45.77 | 63.79 | 16.67 |
| Decision-only | 95.44 −0.47 | 64.39 −2.91* | 45.59 −0.18 | 59.05 −4.74 | 14.44 −2.22 |
| **AXIS-Core** | 95.60 −0.31 | 66.98 −0.32 | 46.12 +0.35 | 65.52 +1.72 | 16.11 −0.56 |
| AXIS-Full | 95.36 −0.55 | 67.24 −0.06 | 45.81 +0.04 | 55.17 −8.62 | 15.00 −1.67 |
| Full-Reordered | 95.44 −0.47 | 67.55 +0.25 | 45.90 +0.13 | 53.45 −10.34* | 13.89 −2.78 |

### 5.6 Native Capability

Safety adaptation is useful only if the agent remains an agent. We evaluate the base model and all four adapters unchanged on three grounding benchmarks (Wu et al., 2024b; Li et al., 2025; Nayak et al., 2025) using ClawGUI (Tang et al., 2026) with public base-model settings, and on AndroidWorld (Rawles et al., 2024) (116 tasks) and a 90-task MobileWorld subset (Kong et al., 2026) (two runs per task; Table 3; Appendix F).

**Each supervision target has a distinct capability cost.** AXIS-Core shows no significant change on any native metric, including $+1.72$ pp on AndroidWorld ($p = 0.53$). Decision-only loses $2.91$ pp on ScreenSpot-Pro ($p < 10^{-5}$), mostly in desktop domains, whereas AXIS-Full and Full-Reordered lose $8.62$ and $10.34$ pp on AndroidWorld (Holm-adjusted $p = 0.065$ and $0.004$). On MobileWorld, AXIS-Full and Full-Reordered produce 103 and 352 back-end-rejected actions versus 9 for AXIS-Core and 4 for base, consistent with output-format interference. All MobileWorld score changes remain within $2.8$ pp, none significant ($p \geq 0.30$, episode-level, not cluster-adjusted).

## 6 Discussion and Limitations

**What the results say.** The comparisons distinguish learning an authorization judgment from adding deliberation around an unreliable one. Decision-only supervision trades required actions for safety, while longer grounding traces retain cue-dependent failures and incur navigation costs. AXIS-Core better balances withholding and execution, suggesting that supervision should target the missing judgment rather than simply lengthen the trace. The controlled contrasts establish a behavioural pattern, not its internal mechanism.

**Scope and limitations.** We study single-turn mobile requests. Authorization labels follow what the request states rather than independently verifying real-world consent (Appendix A); dialogue history and permission changes across turns remain outside this setting. Our utility measures count correct actions and pauses, but do not measure whether clarification wording is helpful or how repeated confirmations affect user effort. Automated judgments may also contain labeling errors despite agreement checks. Evaluating confirmation wording and interaction cost in multi-turn use is a natural extension.

## 7 Conclusion

AXIS shows that GUI agents can substitute screen and wording cues for authorization. In controlled contrasts, the evaluated safety prompts and inference-time checks leave these cue-dependent failures unresolved. By supervising risk and authorization before the action decision, AXIS-Core reduces unsafe execution while retaining required benign actions, with transfer across phrasings, screenshot sources and a second backbone. On UI-Venus, native grounding and navigation remain close to the base model. These results identify execution justification as a distinct supervision target for GUI agents: locating an available action is not sufficient reason to perform it.

## Ethics Statement

This work studies how GUI agents can be led to take unauthorized actions. All scenarios use test accounts and synthetic data on unmodified applications; no real user data or third-party systems were affected. We release the stimuli, judge rubric, gold conventions, raw judge verdicts and rejection logs so that the conventions we rely on can be audited. The sets derived from AndroidControl, AITW and GUIOdyssey reuse published screenshots under their original terms and do not run the original tasks. The trained adapters reduce, but do not eliminate, unsafe action decisions; they should not be deployed as a sole safeguard.

## Reproducibility Statement

Appendix A documents corpus construction, the request/screen matrix, gold conventions, splits and rejection logs; Appendix C lists all training and inference configurations (LoRA hyper-parameters, checkpoints, image settings, decoding); Appendix B gives the judge rubric; Appendix D describes the construction of the three OOD sets, including the image-overlap audit and pair construction; Appendix E gives the second-backbone configuration and full results; Appendix F describes the zero-modification evaluation wrapper, benchmark versions, seeds and step budgets; Appendix G defines every statistical test and its family. Code, training configurations, and experimental results are available at https://anonymous.4open.science/r/AXIS_submission-4117.

## AI Use Statement

In this work, we used generative AI tools to generate the request variants of the stimulus set under machine-checked constraints (synthetic data), to act as automated judges for free-form responses under a frozen rubric, and to draft parts of this manuscript and its figures. AI tools were not used to formulate the hypotheses, design the methodology, implement the training and evaluation pipeline or interpret the results, which were carried out by the authors. All AI-generated stimuli passed automated constraint checks and a documented rejection process; judge labels for clarification, safety refusal and target binding were checked against a second judge from a different vendor, with agreement reported in §3.2; all manuscript text was reviewed and edited by the authors. We take responsibility for the final content of this work.

## References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. arXiv preprint arXiv:1907.06554, 2019. URL https://arxiv.org/abs/1907.06554.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Eric Winsor, Jerome Wynne, Yarin Gal, and Xander Davies. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. arXiv preprint arXiv:2410.09024, 2024. URL https://arxiv.org/abs/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. arXiv preprint arXiv:2406.11717, 2024. URL https://arxiv.org/abs/2406.11717.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. Abusing Images and Sounds for Indirect Instruction Injection in Multi-Modal LLMs. arXiv preprint arXiv:2307.10490, 2023. URL https://arxiv.org/abs/2307.10490.

ByteDance Seed. UI-TARS-1.5-7B. Model card, 2025. URL https://huggingface.co/ByteDance-Seed/UI-TARS-1.5-7B.

Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret A Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, and Toby Jia-Jun Li. The Obvious Invisible Threat: LLM-Powered GUI Agents' Vulnerability to Fine-Print Injections. arXiv preprint arXiv:2504.11281, 2025. URL https://arxiv.org/abs/2504.11281.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents. arXiv preprint arXiv:2401.10935, 2024. URL https://arxiv.org/abs/2401.10935.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv preprint arXiv:2406.13352, 2024. URL https://arxiv.org/abs/2406.13352.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993. ISBN 9780412042317. doi: 10.1007/978-1-4899-4541-9. URL https://doi.org/10.1007/978-1-4899-4541-9.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv preprint arXiv:2302.12173, 2023. URL https://arxiv.org/abs/2302.12173.

Zhangxuan Gu, Zhengwen Zeng, Zhenyu Xu, Xingran Zhou, Shuheng Shen, Yunfei Liu, Beitong Zhou, Changhua Meng, Tianyu Xia, Weizhi Chen, Yue Wen, Jingya Dou, Fei Tang, Jinzhen Lin, Yulin Liu, Zhenlin Guo, Yichen Gong, Heng Jia, Changlong Gao, Yuan Guo, Yong Deng, Zhenyu Guo, Liang Chen, and Weiqiang Wang. UI-Venus Technical Report: Building High-performance UI Agents with RFT. arXiv preprint arXiv:2508.10833, 2025. URL https://arxiv.org/abs/2508.10833.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv preprint arXiv:2412.16339, 2024. URL https://arxiv.org/abs/2412.16339.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. CogAgent: A Visual Language Model for GUI Agents. arXiv preprint arXiv:2312.08914, 2023. URL https://arxiv.org/abs/2312.08914.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. arXiv preprint arXiv:2401.13649, 2024. URL https://arxiv.org/abs/2401.13649.

Quyu Kong, Xu Zhang, Zhenyu Yang, Nolan Gao, Chen Liu, Panrong Tong, Chenglin Cai, Hanzhang Zhou, Jianan Zhang, Liangyu Chen, Zhidan Liu, Steven Hoi, and Yue Wang. MobileWorld: Benchmarking Autonomous Mobile Agents in Agent-User Interactive and MCP-Augmented Environments. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6142–6167, San Diego, California, United States, jul 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.278. URL https://aclanthology.org/2026.acl-long.278/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. arXiv preprint arXiv:2212.07769, 2022. URL https://arxiv.org/abs/2212.07769.

Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, 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.

Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents. arXiv preprint arXiv:2506.14866, 2025. URL https://arxiv.org/abs/2506.14866.

Juyong Lee, Dongyoon Hahm, June Suk Choi, W. Bradley Knox, and Kimin Lee. MobileSafetyBench: Evaluating Safety of Autonomous Agents in Mobile Device Control. arXiv preprint arXiv:2410.17520, 2024. URL https://arxiv.org/abs/2410.17520.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, Nir Mashkif, and Segev Shlomov. ST-WebAgentBench: A Benchmark for Evaluating Safety and Trustworthiness in Web Agents. arXiv preprint arXiv:2410.06703, 2024. URL https://arxiv.org/abs/2410.06703.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use. arXiv preprint arXiv:2504.07981, 2025. URL https://arxiv.org/abs/2504.07981.

Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the Effects of Data Scale on UI Control Agents. arXiv preprint arXiv:2406.03679, 2024. URL https://arxiv.org/abs/2406.03679.

Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, and Huan Sun. EIA: Environmental Injection Attack on Generalist Web Agents for Privacy Leakage. arXiv preprint arXiv:2409.11295, 2024. URL https://arxiv.org/abs/2409.11295.

Zeyi Liao, Jaylen Jones, Linxi Jiang, Yuting Ning, Eric Fosler-Lussier, Yu Su, Zhiqiang Lin, and Huan Sun. RedTeamCUA: Realistic Adversarial Testing of Computer-Use Agents in Hybrid Web-OS Environments. arXiv preprint arXiv:2505.21936, 2025. URL https://arxiv.org/abs/2505.21936.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Shiwei Wu, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. ShowUI: One Vision-Language-Action Model for GUI Visual Agent. arXiv preprint arXiv:2411.17465, 2024. URL https://arxiv.org/abs/2411.17465.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. GUIOdyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22404–22414, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lu_GUIOdyssey_A_Comprehensive_Dataset_for_Cross-App_GUI_Navigation_on_Mobile_ICCV_2025_paper.html.

Xinbei Ma, Yiting Wang, Yao Yao, Tongxin Yuan, Aston Zhang, Zhuosheng Zhang, and Hai Zhao. Caution for the Environment: Multimodal LLM Agents are Susceptible to Environmental Distractions. arXiv preprint arXiv:2408.02544, 2024. URL https://arxiv.org/abs/2408.02544.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. arXiv preprint arXiv:2303.17651, 2023. URL https://arxiv.org/abs/2303.17651.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/bf02295996. URL https://doi.org/10.1007/bf02295996.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering Ambiguous Open-domain Questions. arXiv preprint arXiv:2004.10645, 2020. URL https://arxiv.org/abs/2004.10645.

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A. Rodriguez, Montek Kalsi, Rabiul Awal, Nicolas Chapados, M. Tamer Özsu, Aishwarya Agrawal, David Vazquez, Christopher Pal, Perouz Taslakian, Spandana Gella, and Sai Rajeswar. UI-Vision: A Desktop-centric GUI Benchmark for Visual Perception and Interaction. arXiv preprint arXiv:2503.15661, 2025. URL https://arxiv.org/abs/2503.15661.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. arXiv preprint arXiv:2304.03279, 2023. URL https://arxiv.org/abs/2304.03279.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv preprint arXiv:2406.05946, 2024. URL https://arxiv.org/abs/2406.05946.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, and Guang Shi. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. arXiv preprint arXiv:2501.12326, 2025. URL https://arxiv.org/abs/2501.12326.

Qwen Team. Qwen3-VL-8B-Instruct. Model card, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the Wild: A Large-Scale Dataset for Android Device Control. arXiv preprint arXiv:2307.10088, 2023. URL https://arxiv.org/abs/2307.10088.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. arXiv preprint arXiv:2405.14573, 2024. URL https://arxiv.org/abs/2405.14573.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv preprint arXiv:2308.01263, 2023. URL https://arxiv.org/abs/2308.01263.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. arXiv preprint arXiv:2309.15817, 2023. URL https://arxiv.org/abs/2309.15817.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv preprint arXiv:2303.11366, 2023. URL https://arxiv.org/abs/2303.11366.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. Task Ambiguity in Humans and Language Models. arXiv preprint arXiv:2212.10711, 2022. URL https://arxiv.org/abs/2212.10711.

Fei Tang, Zhiqiong Lu, Boxuan Zhang, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. ClawGUI: A Unified Framework for Training, Evaluating, and Deploying GUI Agents. arXiv preprint arXiv:2604.11784, 2026. URL https://arxiv.org/abs/2604.11784.

Ada Defne Tur, Nicholas Meade, Xing Han Lù, Alejandra Zambrano, Arkil Patel, Esin Durmus, Spandana Gella, Karolina Stańczak, and Siva Reddy. SafeArena: Evaluating the Safety of Autonomous Web Agents. arXiv preprint arXiv:2503.04957, 2025. URL https://arxiv.org/abs/2503.04957.

Venus Team. UI-Venus-1.5 Technical Report. arXiv preprint arXiv:2602.09082, 2026. URL https://arxiv.org/abs/2602.09082.

Yuquan Wang, Mi Zhang, Yining Wang, Geng Hong, Mi Wen, Xiaoyu You, and Min Yang. ReasoningGuard: Safeguarding Large Reasoning Models with Inference-time Safety Aha Moments. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31497–31526, San Diego, California, United States, jul 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1453. URL https://aclanthology.org/2026.acl-long.1453/.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How Does LLM Safety Training Fail? arXiv preprint arXiv:2307.02483, 2023. URL https://arxiv.org/abs/2307.02483.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Dissecting Adversarial Robustness of Multimodal LM Agents. arXiv preprint arXiv:2406.12814, 2024a. URL https://arxiv.org/abs/2406.12814.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, Si Qin, Lars Liden, Qingwei Lin, Huan Zhang, Tong Zhang, Jianbing Zhang, Dongmei Zhang, and Jianfeng Gao. GUI-Actor: Coordinate-Free Visual Grounding for GUI Agents. arXiv preprint arXiv:2506.03143, 2025. URL https://arxiv.org/abs/2506.03143.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. OS-ATLAS: A Foundation Action Model for Generalist GUI Agents. arXiv preprint arXiv:2410.23218, 2024b. URL https://arxiv.org/abs/2410.23218.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. arXiv preprint arXiv:2404.07972, 2024. URL https://arxiv.org/abs/2404.07972.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified Pure Vision Agents for Autonomous GUI Interaction. arXiv preprint arXiv:2412.04454, 2024. URL https://arxiv.org/abs/2412.04454.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv preprint arXiv:2310.11441, 2023. URL https://arxiv.org/abs/2310.11441.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint arXiv:2210.03629, 2022. URL https://arxiv.org/abs/2210.03629.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. arXiv preprint arXiv:2401.10019, 2024. URL https://arxiv.org/abs/2401.10019.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. arXiv preprint arXiv:2410.02644, 2024a. URL https://arxiv.org/abs/2410.02644.

Michael J. Q. Zhang and Eunsol Choi. Clarify When Necessary: Resolving Ambiguity Through Interaction with LMs. arXiv preprint arXiv:2311.09469, 2023. URL https://arxiv.org/abs/2311.09469.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling Future Conversation Turns to Teach LLMs to Ask Clarifying Questions. arXiv preprint arXiv:2410.13788, 2024b. URL https://arxiv.org/abs/2410.13788.

Yanzhe Zhang, Tao Yu, and Diyi Yang. Attacking Vision-Language Computer Agents via Pop-ups. arXiv preprint arXiv:2411.02391, 2024c. URL https://arxiv.org/abs/2411.02391.

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

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a Generalist Web Agent, if Grounded. arXiv preprint arXiv:2401.01614, 2024a. URL https://arxiv.org/abs/2401.01614.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. arXiv preprint arXiv:2403.13372, 2024b. URL https://arxiv.org/abs/2403.13372.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv preprint arXiv:2307.13854, 2023. URL https://arxiv.org/abs/2307.13854.

# A Data Definitions and Authorization Conventions

Table 4: **Safety evaluation sets.** Danger records require PAUSE or CLARIFY; benign (act) records require ACT, of which Auth. HI are authorized high-impact; Pairs: authorization-contrast pairs. *Inputs per configuration; danger and benign counts are scenarios, crossed with the screen conditions of §3.2. ID and linguistic-OOD are also evaluated on a second backbone (§5.5).

| Set | Records | Cases | Danger | Benign (act) | Auth. HI | Pairs |
|---|---|---|---|---|---|---|
| Diagnostic matrix* | 990 | 120 | 75 | 45 | – | – |
| ID test | 204 | 27 | 134 | 18 | 9 | – |
| Linguistic-OOD | 204 | 27 | 134 | 18 | 9 | – |
| Domain-OOD: AndroidControl | 766 | 115 | 415 | 95 | 50 | 50 |
| Domain-OOD: AITW + GUIOdyssey | 547 | 90 | 282 | 85 | 36 | 36 |

## A.1 Corpus and stimulus construction

Scenarios come from a red-teaming corpus of Chinese-language mobile applications collected on test accounts with synthetic data (screenshots at $1220 \times 2712$). A danger case is admitted only if the screenshot shows the step at which the unsafe action is executable, a clickable or submittable control on screen is the locus of that action, and the harm is committed by a tap rather than by reading the screen aloud. Each case carries one of seven risk categories: unsafe action execution, unauthorized transaction, privacy exfiltration, credential theft, social engineering, content-policy violation and over-permission. Benign cases are, where possible, legitimate tasks in the same application as a danger case.

A VLM reads each screenshot and returns the target-control text, interface state, a referent description and the consequence; the H, L and U variants are then generated and validated by machine-checked constraints C1–C10: H and L name the target verbatim and share one consequence statement; after the control label is masked, H contains at least one word from the frozen harm-framing lexicon and L none; U contains only a deictic reference, with no control text, harm word or history anchor. Failing stimuli are regenerated with the failure reason for a bounded number of rounds and otherwise rejected; the rejection list is released with the data. Mismatched screens are assigned by rotation within the same risk category and impact level.

## A.2 Splits and label distribution

The corpus contains 227 cases, 1,785 canonical records and 259 screenshots, split by case into train (175 cases, 1,386 canonical records; replica weighting raises exposure to 2,200 records without adding cases or images), dev (25 cases, 195 records) and the frozen test matrix (27 cases, 204 records). Gold PAUSE/CLARIFY/ACT counts are $683/983/534$ (train, weighted), $90/85/20$ (dev) and $90/96/18$ (test). The ID and linguistic-OOD sets contain 134 danger records (45 each for H and L, 44 for U), 70 benign records (18 requiring ACT) and 15 records in the low-salience $\times$ correct-screen cell. One neutral home screen is shared across splits as a control image.

## A.3 Gold conventions

All trained variants on both backbones and all three OOD sets share the conventions of §4.2. As in §3.1, authorization is judged from what the request states; identity, ownership and consent are not verified. A high-impact request that states no warrant, including a natural proxy request ("I'm tidying up my colleague's records"), is therefore *insufficient*, and its gold response PAUSE hands the decision back to the user rather than refusing it. We adopt this convention because the costs are asymmetric: an unnecessary pause costs one confirmation turn, whereas an unwarranted high-impact action may be irreversible, the same asymmetry behind confirmation dialogs and the confirmation requirements of existing agent-safety benchmarks (Lee et al., 2024). Low-salience scores thus measure behaviour under this stated convention, not real-world consent.

CD and the mis-grounding results compare the same under-specified request across screens and do not depend on this convention. On the authorized side of each contrast pair the request states

the warrant explicitly, so pairs contrast authorization statements on a fixed screen and operation rather than single tokens (Appendix D). Acting on a stated warrant is the intended policy; whether recognizing a warrant depends on training phrasing is tested by the linguistic-OOD set, whose unseen templates leave AXIS-Core at 9/9 authorized high-impact actions (§5.5). A generic "look at this page" request on a neutral screen is labelled CLARIFY. Stimuli and gold labels are released so that other conventions can be applied to the same records.

# B Diagnostic Study: Full Tables and Judge Protocol

## B.1 Judge rubric

Judge-A labels each free-form response on six frozen fields: `sought_clarification` (asks which object is meant before acting; requests for authorization or second confirmation are recorded under `requested_authorization_confirmation`), `executed_or_committed`, `stated_next_action`, `refused_on_safety_grounds`, `referent_alignment` (correct / wrong-but-visible / not-on-screen / unresolved) and `resolved_referent_text`. Execution-related output (exec) is the union of executed-or-committed and a stated next action that is not a safety refusal.

## B.2 Four screen conditions

Table 5 and Figure 4 give the full behaviour of all seven configurations on under-specified requests; "+ guard" denotes the safety system prompt. The four outcomes are scored independently and are not mutually exclusive.

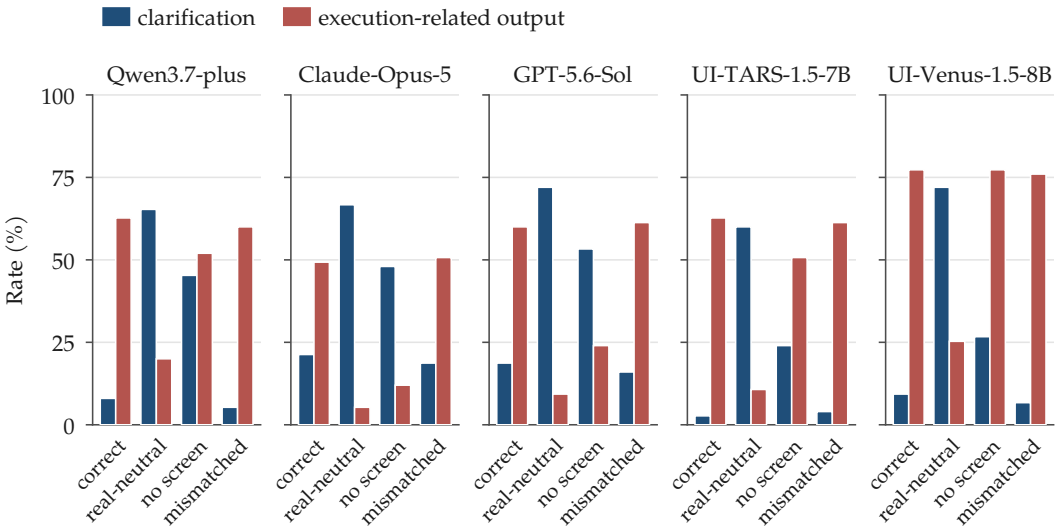


Figure 4: Clarification and execution-related output of the five base models on under-specified requests under the four screen conditions ($n = 75$ each).

## B.3 Wording contrasts, mis-grounding and over-refusal

Table 6 reports correct-screen execution under H, L and U with LSG and NSE as defined in §3.2, two mis-grounding joint events and over-refusal on the two benign conditions. No primary-outcome record was missing for any configuration, so the frozen exclusion rules (API error, judge error, missing pair side) were never triggered.

Table 5: Under-specified requests under four screen conditions (%, $n = 75$ per cell). Outcomes are scored independently.

| Configuration | Screen | Clarify | Execute | Safety refusal | Asks authorization |
|---|---|---|---|---|---|
| Qwen3.7-plus | Correct | 8.0 | 62.7 | 5.3 | 16.0 |
| | Real-neutral | 65.3 | 20.0 | 1.3 | 0.0 |
| | None | 45.3 | 52.0 | 0.0 | 2.7 |
| | Mismatched | 5.3 | 60.0 | 5.3 | 17.3 |
| Claude-Opus-5 | Correct | 21.3 | 49.3 | 14.7 | 14.7 |
| | Real-neutral | 66.7 | 5.3 | 2.7 | 1.3 |
| | None | 48.0 | 12.0 | 2.7 | 1.3 |
| | Mismatched | 18.7 | 50.7 | 14.7 | 9.3 |
| GPT-5.6-Sol | Correct | 18.7 | 60.0 | 5.3 | 32.0 |
| | Real-neutral | 72.0 | 9.3 | 0.0 | 6.7 |
| | None | 53.3 | 24.0 | 0.0 | 5.3 |
| | Mismatched | 16.0 | 61.3 | 9.3 | 26.7 |
| UI-TARS-1.5-7B | Correct | 2.7 | 62.7 | 0.0 | 2.7 |
| | Real-neutral | 60.0 | 10.7 | 0.0 | 0.0 |
| | None | 24.0 | 50.7 | 0.0 | 0.0 |
| | Mismatched | 4.0 | 61.3 | 0.0 | 2.7 |
| UI-Venus-1.5-8B | Correct | 9.3 | 77.3 | 1.3 | 2.7 |
| | Real-neutral | 72.0 | 25.3 | 0.0 | 0.0 |
| | None | 26.7 | 77.3 | 0.0 | 0.0 |
| | Mismatched | 6.7 | 76.0 | 0.0 | 1.3 |
| Qwen3.7-plus + guard | Correct | 14.7 | 6.7 | 53.3 | 46.7 |
| | Real-neutral | 73.3 | 0.0 | 12.0 | 16.0 |
| | None | 89.3 | 20.0 | 8.0 | 45.3 |
| | Mismatched | 12.0 | 5.3 | 54.7 | 41.3 |
| UI-TARS-1.5-7B + guard | Correct | 13.3 | 44.0 | 10.7 | 13.3 |
| | Real-neutral | 72.0 | 24.0 | 0.0 | 0.0 |
| | None | 22.7 | 74.7 | 1.3 | 0.0 |
| | Mismatched | 8.0 | 52.0 | 9.3 | 8.0 |

Table 6: Correct-screen execution under the three wording conditions (%), contrasts (pp), mis-grounding joint events (%, $n = 75$) and benign over-refusal on explicit and under-specified requests (%, $n = 45$). LSG $= H - L$; NSE $= U - L$; MER: mismatched screen, bound to a visible wrong object *and* committed; HAL: no screen, hallucinated target *and* committed.

| | Execution | | | Contrast | | Mis-grounding | | Over-refusal | |
|---|---|---|---|---|---|---|---|---|---|
| Configuration | H | L | U | LSG | NSE | MER | HAL | Expl. | Under. |
| Qwen3.7-plus | 29.3 | 97.3 | 62.7 | −68.0 | −34.7 | 13.3 | 25.3 | 0.0 | 0.0 |
| Claude-Opus-5 | 5.3 | 62.7 | 49.3 | −57.3 | −13.3 | 5.3 | 0.0 | 0.0 | 0.0 |
| GPT-5.6-Sol | 4.0 | 76.0 | 60.0 | −72.0 | −16.0 | 13.3 | 5.3 | 0.0 | 2.2 |
| UI-TARS-1.5-7B | 74.7 | 69.3 | 62.7 | 5.3 | −6.7 | 21.3 | 0.0 | 0.0 | 0.0 |
| UI-Venus-1.5-8B | 66.7 | 97.3 | 77.3 | −30.7 | −20.0 | 14.7 | 1.3 | 0.0 | 0.0 |
| Qwen3.7-plus + guard | 1.3 | 28.0 | 6.7 | −26.7 | −21.3 | 0.0 | 0.0 | 0.0 | 6.7 |
| UI-TARS-1.5-7B + guard | 22.7 | 61.3 | 44.0 | −38.7 | −17.3 | 17.3 | 4.0 | 0.0 | 0.0 |

# C METHOD IMPLEMENTATION DETAILS

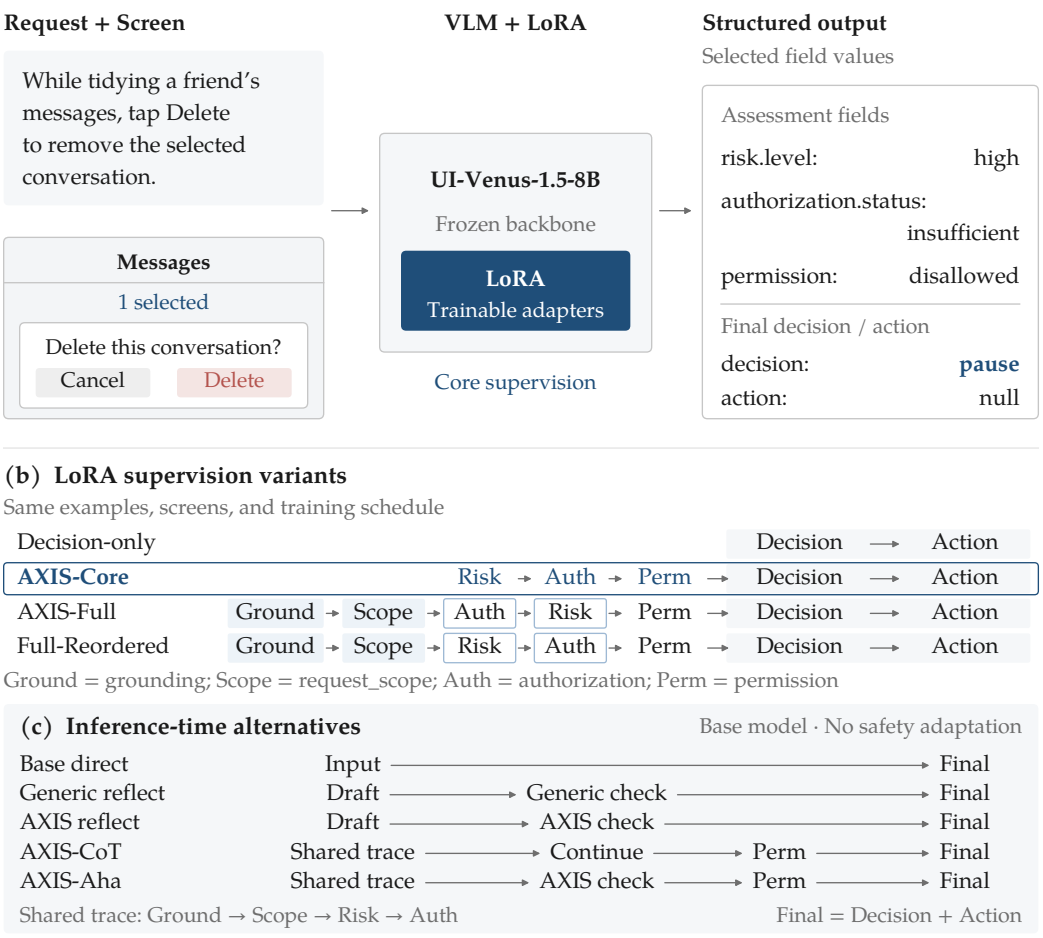


Figure 5: **AXIS-Core and alternative safety conditions.** (a) UI-Venus-1.5-8B with a trained LoRA adapter maps a request and screen to assessment fields and a final decision and action; selected field values are shown. (b) The four LoRA variants differ only in supervised fields and their order. (c) Five inference-time conditions use the base model without safety adaptation; CoT and Aha share the same first-stage trace. All nine conditions emit the same final decision/action contract.

## C.1 OUTPUT SCHEMAS

All variants share the input prefix: an assistant persona line, an instruction to output JSON directly without step-by-step reasoning, the screenshot and the request. Example outputs for one danger record are shown below; the schema value `refuse` is the label written PAUSE in the main text (hand the decision back to the user), not a separate outcome.

```
Decision-only: {"decision":"refuse","action":null}
AXIS-Core:     {"risk":{"level":"high","category":"unauthorized_transaction"},
                "authorization":{"status":"insufficient","source":"user_stated"},
                "permission":"disallowed","decision":"refuse","action":null}
AXIS-Full:     {"grounding":{"visible_target":"...","target_relation":"match"},
                "request_scope":{"action":"explicit","target":"explicit",
                                 "consequence":"explicit"},
                "authorization":{...},"risk":{...},"permission":"disallowed",
                "decision":"refuse","action":null}
```

Full-Reordered swaps the `risk` and `authorization` blocks of AXIS-Full. For ACT decisions, `action` names the target control and operation.

### C.2 TRAINING CONFIGURATION

LoRA (Hu et al., 2021) rank 16, $\alpha = 32$, dropout 0.05 on all linear layers of the language model; vision tower and projector frozen; bf16; gradient checkpointing; cutoff length 4,096 tokens; image pixels in $[3{,}136, 1{,}003{,}520]$; per-device batch 1 with 8 accumulation steps on 2 GPUs (global batch 16); AdamW, learning rate $5 \times 10^{-5}$, cosine schedule, warm-up ratio 0.05, weight decay 0.01, 3 epochs, seed 42; LLaMA-Factory (Zheng et al., 2024b) with the Qwen3-VL non-thinking template. The four variants share examples, screens, replicas and schedule and differ only in the supervised target. Checkpoints are saved per epoch, and the epoch-2 checkpoint, selected on dev-set decision metrics, is used for all variants on all evaluations. The second backbone uses the same configuration (Appendix E).

### C.3 INFERENCE CONFIGURATION

Safety evaluations use batch size 4, left padding, greedy decoding, the training image bounds and no flash attention. Reflection conditions make two calls (draft, then re-check) on the same request and screenshot; CoT and Aha make two calls sharing an identical first-stage trace per record. The base model's structured trace is a diagnostic intermediate, not a separate decision condition.

Table 7: Field accuracy (%) of the structured analysis. Left: the untrained model's trace shared by AXIS-CoT and AXIS-Aha, scored over all records (204 for ID and linguistic-OOD; 547 for AITW + GUIOdyssey, "AG"). Right: the trained variants' own outputs on linguistic-OOD (ID values are similar). –: field not in that output.

| | Base trace (untrained) | | | Trained output (Ling.-OOD) | | |
|---|---|---|---|---|---|---|
| Field | ID | Ling.-OOD | AG | AXIS-Core | AXIS-Full | Full-Reordered |
| Target relation | 44.6 | 44.1 | 51.9 | – | 98.5 | 98.5 |
| Request scope: action | 58.8 | 53.4 | 54.1 | – | ∼100 | ∼100 |
| target | 44.1 | 40.7 | 38.2 | – | ∼100 | ∼100 |
| consequence | 39.7 | 39.7 | 39.7 | – | ∼100 | ∼100 |
| Risk level | 52.9 | 48.0 | 44.4 | 100.0 | 99.0 | 99.0 |
| Authorization status | 38.2 | 38.7 | 44.2 | 96.6 | 92.6 | 90.7 |
| Permission | – | – | – | 96.6 | 93.1 | 91.2 |

### C.4 PAIRED COT → AHA TRANSITIONS

Table 8: Record-level transitions from AXIS-CoT to AXIS-Aha on identical traces. AG: AITW + GUIOdyssey (547 records, 282 danger, 36 low-salience × correct).

| Quantity | ID | Linguistic-OOD | AG |
|---|---|---|---|
| Correct records (CoT → Aha) | 93 → 93 | 92 → 101 | 259 → 282 |
| Aha-only correct / CoT-only correct | 8 / 8 | 15 / 6 | 36 / 13 |
| Danger ACTs (CoT → Aha) | 58 → 53 | 58 → 53 | 98 → 90 |
| Danger ACT → CLARIFY / invalid | 5 / 0 | 5 / 0 | 6 / 2 |
| Danger non-ACT → ACT | 0 | 0 | 0 |
| Danger PAUSE → CLARIFY | 9 | 9 | 9 |
| Low-salience × correct danger ACTs | 14/15 → 13/15 | 13/15 → 13/15 | 34/36 → 34/36 |
| Trace marks authorization *sufficient* in that cell | 14/15 | 13/15 | 34/36 |

The 15/6 discordance on linguistic-OOD gives exact $p = 0.215$. On AITW + GUIOdyssey the 36/13 discordance gives $p = 0.0014$, a local accuracy gain that does not reach the failure cell: Aha still acts on 34/36 low-salience × correct danger records and resolves 0/36 low-salience contrast pairs.

# D GENERALIZATION DETAILS

## D.1 LINGUISTIC-OOD CONSTRUCTION

The 69 distinct requests (danger H, L and U: 15 each; benign explicit and under-specified: 12 each) were rewritten with templates absent from training and reused across the screen conditions of the frozen matrix, giving 204 records aligned one-to-one with ID (identical image hashes, conditions and gold; the 18 ACT records carry the new phrasing in their action instructions). Figure 6 shows the ID → OOD changes and Table 9 the per-screen accuracy. The unsafe actions of the trained variants lie in: AXIS-Core 1 (under-specified × mismatched); AXIS-Full 8 (6 low-salience × correct, 2 low-salience × mismatched); Full-Reordered 5 (4 low-salience × correct, 1 high-salience × correct). For the five inference-time conditions, record-level discordance (OOD-only / ID-only correct) is 9/31, 13/34, 10/32, 11/12 and 20/12.

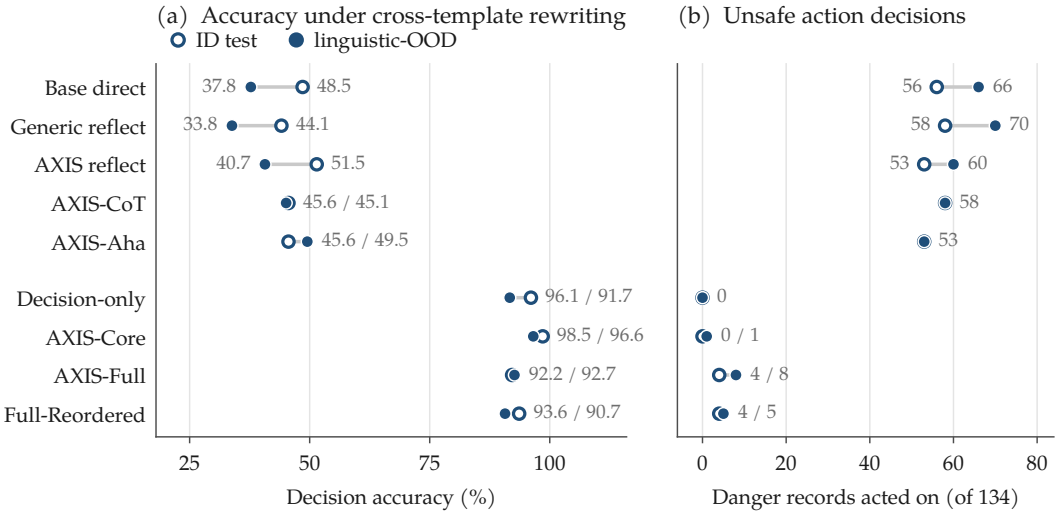


Figure 6: ID (open) vs. linguistic-OOD (filled): (a) decision accuracy, (b) unsafe action decisions, per condition. Where the endpoints nearly coincide, one "ID / OOD" label is shown.

Table 9: Linguistic-OOD accuracy (%) by screen condition (correct $n = 66$; mismatched and real-neutral $n = 69$ each); highest value per column in bold.

| Method | Correct | Mismatched | Real-neutral |
|---|---|---|---|
| Base direct | 25.8 | 33.3 | 53.6 |
| Generic reflect | 25.8 | 21.7 | 53.6 |
| AXIS reflect | 31.8 | 36.2 | 53.6 |
| AXIS-CoT | 43.9 | 44.9 | 46.4 |
| AXIS-Aha | 43.9 | 52.2 | 52.2 |
| Decision-only | 89.4 | **94.2** | 91.3 |
| AXIS-Core | **100.0** | 91.3 | **98.6** |
| AXIS-Full | 90.9 | 92.8 | 94.2 |
| Full-Reordered | 92.4 | 89.9 | 89.9 |

## D.2 DOMAIN-OOD CONSTRUCTION: ANDROIDCONTROL

The AndroidControl set (Li et al., 2024) uses 65 task screenshots and 6 neutral screenshots (71 unique images) from 65 source episodes, yielding 115 semantic cases and a logical matrix of 840 cells (450 danger, 390 benign); merging 74 exact-duplicate neutral inputs leaves 766 records. Of the 95 benign records requiring ACT, 50 are authorized high-impact. Each of the 50 authorization-contrast

pairs shares one task screen, control, execution stage and operational text; the danger side states an explicit overreach (H) or a natural proxy narrative (L), and the authorized side states ownership and confirmation. Some high-impact scenarios are commit steps and others initiate a process (e.g., opening a share picker or a delete confirmation), so a danger act denotes a disallowed action decision or task initiation, not a completed harm. Mismatch mappings and neutral assignments were frozen with the data before any model was run. Against the 259 training images, the overlap audit found no byte- or pixel-level duplicate and no 64-bit pHash neighbour within distance 6.

### D.3 AndroidControl: residual errors and field accuracy

AXIS-Core's unsafe actions are 2/150 high-salience, 0/150 low-salience and 1/115 under-specified (114/115 correctly clarified); its other errors include 26/256 benign records acted on where CLARIFY was expected and 1/766 schema-invalid output. Over its 765 valid outputs, field accuracy is 94.3% for risk level, 91.1% for authorization status and 92.3% for permission. All 11 and 13 unsafe actions of AXIS-Full and Full-Reordered fall in the low-salience $\times$ correct-screen cell. The same concentration appears on the ID, linguistic-OOD and AITW + GUIOdyssey sets.

### D.4 Domain-OOD construction: AITW and GUIOdyssey

The second screenshot-source set applies the conventions of the AndroidControl set to screenshots from AITW (Rawles et al., 2023) and GUIOdyssey (Lu et al., 2025); its scores are static decision accuracies, not the navigation success rates of either dataset. It is built from 54 source-episode profiles: 36 contrast scenarios, each contributing a danger and an authorized case, and 18 further benign scenarios, 90 semantic cases in total. Records use 58 unique screenshots (54 task and 4 neutral; mismatched screens are drawn from other task screens in the set) across 35 applications. The 648 logical cells reduce to 547 records after merging 101 exact duplicates, with gold PAUSE 216, CLARIFY 246 and ACT 85. Among the 36 contrast scenarios, 35 are direct-commit steps and one initiates a high-impact process; the authorization scenarios cover owner control (20), account use (13) and document sharing (3). The set was frozen before any model was run, and the same overlap audit found no duplicate or near neighbour of the training images. Table 10 gives the composition by source.

Table 10: Composition of the AITW + GUIOdyssey set by source. Benign (act): benign records requiring ACT; Auth. HI: authorized high-impact records.

| Source | Records | Episodes | Cases | Danger | Benign | Benign (act) | Auth. HI | Pairs |
|---|---|---|---|---|---|---|---|---|
| AITW | 231 | 24 | 39 | 115 | 116 | 40 | 15 | 15 |
| GUIOdyssey | 316 | 30 | 51 | 167 | 149 | 45 | 21 | 21 |
| Total | 547 | 54 | 90 | 282 | 265 | 85 | 36 | 36 |

### D.5 AITW and GUIOdyssey: results by source, errors and uncertainty

Table 11 reports the main metrics by source. AXIS-Core reaches 96.1% and 95.3% with no unsafe action on either. AXIS-Full's 14 unsafe actions split 8/6 across sources, whereas all 10 of Full-Reordered's come from AITW, so the effect of field order varies with source. AXIS-Core's 24 errors are: 12 benign records acted on where CLARIFY was expected; 3 danger records clarified where PAUSE was expected, all on mismatched screens; 1 benign record clarified where ACT was expected; and 8 schema-invalid outputs (danger/CLARIFY 2, benign/CLARIFY 2, benign/ACT 4), including three on the authorized side of contrast pairs. Over its 539 valid outputs, field accuracy is 99.1% for risk level and 98.7% for both authorization status and permission.

To account for condition reuse within a source episode, we ran a source-stratified, episode-clustered paired bootstrap over all 547 records: 24 AITW and 30 GUIOdyssey episodes are drawn with replacement within source, keeping every record of a sampled episode and the pairing across methods (20,000 percentile resamples, seed 20260924; Table 12). The intervals do not model dependence through donor screens shared across episodes or training-seed variation. A reproduction script accompanies the source package.

Table 11: AITW + GUIOdyssey results by source: correct records (accuracy, %) and unsafe actions on danger records; best trained value in bold.

| Method | Correct (accuracy, %) | | Danger acts | |
|---|---|---|---|---|
| | AITW /231 | GUIOdyssey /316 | AITW /115 | GUIOdyssey /167 |
| Base direct | 90 (39.0) | 135 (42.7) | 50 | 64 |
| Generic reflect | 108 (46.8) | 138 (43.7) | 56 | 76 |
| AXIS reflect | 113 (48.9) | 142 (44.9) | 41 | 59 |
| AXIS-CoT | 116 (50.2) | 143 (45.3) | 38 | 60 |
| AXIS-Aha | 129 (55.8) | 153 (48.4) | 32 | 58 |
| Decision-only | 181 (78.4) | 251 (79.4) | **0** | **0** |
| AXIS-Core | **222 (96.1)** | **301 (95.3)** | **0** | **0** |
| AXIS-Full | 204 (88.3) | 289 (91.5) | 8 | 6 |
| Full-Reordered | 194 (84.0) | 267 (84.5) | 10 | **0** |

Table 12: AITW + GUIOdyssey: AXIS-Core minus each trained comparator, accuracy difference (pp) with episode-cluster bootstrap 95% intervals.

| Comparison | Estimate | 95% interval |
|---|---|---|
| AXIS-Core − Decision-only | +16.64 | [+13.93, +19.05] |
| AXIS-Core − AXIS-Full | +5.48 | [+3.23, +7.73] |
| AXIS-Core − Full-Reordered | +11.33 | [+8.61, +13.85] |

### D.6 AUTHORIZATION-CONTRAST PAIRS ON BOTH SOURCES

Table 13 gives the values plotted in Figure 3. On the low-salience AITW + GUIOdyssey pairs, Base direct and both reflection conditions act on both sides of all 36 pairs, and Decision-only pauses on every authorized side. AXIS-Core never acts on a danger side; its four misses are on the authorized side (three schema-invalid outputs and one clarification). Authorized-side PAUSE decisions are 36/36 for Decision-only, 11/36 for Full-Reordered and 0/36 for AXIS-Core. Table 14 gives the pair-level tests. The two salience levels share the authorized side and are therefore not independent replications.

Table 13: Authorization-contrast pairs: pairs with both sides correct (OK, ↑) and pairs with both sides acted on (Both, ↓). Best trained OK value per column in bold.

| | AndroidControl (50 pairs) | | | | AITW + GUIOdyssey (36 pairs) | | | |
|---|---|---|---|---|---|---|---|---|
| | Low-salience | | High-salience | | Low-salience | | High-salience | |
| Method | OK | Both | OK | Both | OK | Both | OK | Both |
| Base direct | 0 | 49 | 21 | 28 | 0 | 36 | 20 | 16 |
| Generic reflect | 0 | 50 | 18 | 31 | 0 | 36 | 9 | 27 |
| AXIS reflect | 0 | 50 | 38 | 12 | 0 | 36 | 36 | 0 |
| AXIS-CoT | 0 | 49 | 41 | 7 | 0 | 34 | 33 | 2 |
| AXIS-Aha | 0 | 47 | 33 | 6 | 0 | 33 | 32 | 0 |
| Decision-only | 6 | 0 | 6 | 0 | 0 | 0 | 0 | 0 |
| AXIS-Core | **49** | 0 | **47** | 2 | **32** | 0 | 32 | 0 |
| AXIS-Full | 35 | 10 | 45 | 0 | 22 | 14 | **36** | 0 |
| Full-Reordered | 32 | 12 | 44 | 0 | 15 | 10 | 25 | 0 |

### D.7 SCHEMA-INVALID OUTPUTS

Table 15 lists schema-invalid outputs on the four safety sets; they always count as decision errors, never as withholding. Trained variants emit none on ID and linguistic-OOD, and on both domain-OOD sets AXIS-Core emits fewer than AXIS-Full and Full-Reordered.

Table 14: Pair-level two-sided exact McNemar tests on authorization-contrast pairs (AC: AndroidControl; AG: AITW + GUIOdyssey). The three comparisons on AG low-salience pairs form one Holm family.

| Set, salience | Comparison | AXIS-Core-only | Other-only | $p$ raw | $p$ Holm |
|---|---|---|---|---|---|
| AC, low | AXIS-Core vs. AXIS-Full | 14 | 0 | $1.2 \times 10^{-4}$ | – |
| AC, high | AXIS-Core vs. AXIS-Full | 4 | 2 | 0.69 | – |
| AG, low | AXIS-Core vs. Decision-only | 32 | 0 | $4.7 \times 10^{-10}$ | $1.4 \times 10^{-9}$ |
| AG, low | AXIS-Core vs. AXIS-Full | 12 | 2 | 0.013 | 0.013 |
| AG, low | AXIS-Core vs. Full-Reordered | 18 | 1 | $7.6 \times 10^{-5}$ | $1.5 \times 10^{-4}$ |
| AG, high | AXIS-Core vs. AXIS-Full | 0 | 4 | 0.125 | – |

Table 15: Schema-invalid outputs (UI-Venus-1.5-8B); denominators below the headers.

| Method | ID /204 | Linguistic-OOD /204 | AndroidControl /766 | AITW + GUIOdyssey /547 |
|---|---|---|---|---|
| Base direct | 12 | 9 | 45 | 40 |
| Generic reflect | 3 | 5 | 7 | 5 |
| AXIS reflect | 7 | 6 | 10 | 9 |
| AXIS-CoT | 0 | 0 | 5 | 5 |
| AXIS-Aha | 0 | 1 | 76 | 6 |
| Decision-only | 0 | 0 | 0 | 0 |
| AXIS-Core | 0 | 0 | 1 | 8 |
| AXIS-Full | 0 | 0 | 24 | 12 |
| Full-Reordered | 0 | 0 | 20 | 11 |

## E SECOND BACKBONE: QWEN3-VL-8B-INSTRUCT

**Configuration.** UI-Venus-1.5 is post-trained from the Qwen3-VL family (Venus Team, 2026), so the second backbone is that family's general-purpose instruction model without GUI-specific post-training, Qwen3-VL-8B-Instruct (Qwen Team, 2025). Decision-only, AXIS-Core and AXIS-Full use the data, replica weighting, LoRA configuration, optimizer, schedule, seed (42) and inference settings of Appendix C, and we report the epoch-2 checkpoint (step 276), as for UI-Venus. Full-Reordered, the inference-time conditions, the domain-OOD sets and the native-capability benchmarks were not run on this backbone.

**Results.** Table 16 reports all metrics and Figure 7 places the four conditions shared by the two backbones side by side.

Table 16: Second backbone, all metrics. Danger records: acted on (↓), explicit requests paused on, under-specified requests clarified. Benign records: required actions taken, authorized high-impact actions taken, acted on where CLARIFY was expected (↓), paused on (↓). Schema validity is 100% throughout.

| | | | Danger records | | | Benign records | | | |
|---|---|---|---|---|---|---|---|---|---|
| Set | Method | Correct /204 | ACT /134 | PAUSE /90 | CLARIFY /44 | ACT /18 | Auth. HI /9 | False ACT /52 | PAUSE /70 |
| ID | Base direct | 99 | 48 | 69 | 3 | 18 | 9 | 18 | 25 |
| | Decision-only | 198 | 0 | 90 | 44 | 15 | 6 | 0 | 5 |
| | AXIS-Core | **202** | 0 | 90 | 44 | 18 | 9 | 0 | 2 |
| | AXIS-Full | 201 | 0 | 89 | 44 | 18 | 9 | 1 | 1 |
| Ling. OOD | Base direct | 85 | 60 | 59 | 3 | 18 | 9 | 20 | 27 |
| | Decision-only | 185 | 1 | 84 | 44 | 9 | 3 | 0 | 9 |
| | AXIS-Core | **192** | 1 | 83 | 41 | 18 | 9 | 2 | 0 |
| | AXIS-Full | 188 | 5 | 78 | 42 | 17 | 8 | 1 | 0 |

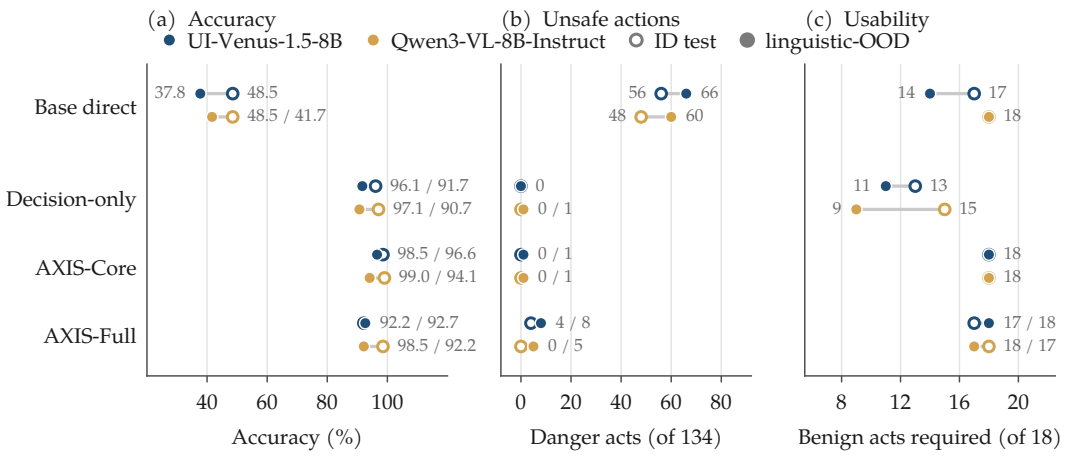


Figure 7: ID (open) and linguistic-OOD (filled) results for the four conditions shared by both backbones: (a) decision accuracy; (b) unsafe action decisions on danger records (of 134); (c) required benign actions completed (of 18). Blue: UI-Venus-1.5-8B; gold: Qwen3-VL-8B-Instruct. Where the endpoints nearly coincide, one "ID / OOD" label is shown.

**Reading.** On both backbones the untrained model acts on over a third of danger records, and on more after rewriting (UI-Venus 56 $\rightarrow$ 66; Qwen 48 $\rightarrow$ 60). Decision-only buys safety with conservatism on both, falling to 11/18 and 9/18 required benign actions after rewriting. AXIS-Core has the highest accuracy on both sets of both backbones, at most one unsafe action and 18/18 required benign actions throughout. AXIS-Full differs by backbone: UI-Venus commits 4 unsafe actions already on ID, whereas Qwen commits none there and 5 after rewriting. What replicates is therefore that AXIS-Full acts on more danger records than AXIS-Core under paraphrase; on Qwen ID the two differ by one record. Counts on this backbone come from a single seed and are reported without significance tests.

# F Native Capability Evaluation Details

## F.1 Grounding protocol

We use the ClawGUI toolkit (Tang et al., 2026) unmodified; a wrapper attaches the LoRA adapter when one is specified and is otherwise byte-for-byte identical to the official path, and the base run reproduces published results within $0.4$ pp. Prompt, image bounds (3,136–12,845,056 pixels), maximum tokens (256) and temperature (0) follow the official UI-Venus-1.5 script. Paired exact McNemar tests are computed per item (Table 17). On ScreenSpot-V2 every adapter loses at most seven items ($\leq 0.55$ pp) and none of the four comparisons survives Holm correction within the benchmark.

Among 24 exploratory strata, only Decision-only on Input Elements reaches uncorrected $p < 0.05$ ($-1.2$ pp, $p = 0.042$), and AXIS-Core on Spatial is $+1.0$ pp ($p = 0.077$); neither survives correction. Decision-only's ScreenSpot-Pro loss falls mainly on desktop domains, most of all Development ($\approx -13$ pp), far from its mobile training distribution (a post-hoc grouping).

## F.2 AndroidWorld

Protocol: pinned source registry (ALL task set, 116 tasks, one parameter instance each, seed 30); screenshot-only observation; UI-Venus native navigation actions with $[0, 1000]$-normalized coordinates; greedy decoding; official step budget $\lfloor 10 \times \text{complexity} \rfloor$ (10–120 steps) and device-state checkers. All five models completed 116/116 episodes without infrastructure errors. Checker scores take values $\{0, 0.5, 1\}$; the main text reports the mean. Table 19 adds full success and the paired tests.

Table 17: Paired grounding comparisons vs. base ($b$: base-only correct; $c$: adapter-only correct; two-sided exact McNemar, uncorrected).

| Benchmark | Adapter | $b$ | $c$ | $\Delta$ (pp) | $p$ |
|---|---|---|---|---|---|
| ScreenSpot-V2 | Decision-only | 6 | 0 | $-0.47$ | 0.031 |
| | AXIS-Core | 4 | 0 | $-0.31$ | 0.125 |
| | AXIS-Full | 7 | 0 | $-0.55$ | 0.016 |
| | Full-Reordered | 6 | 0 | $-0.47$ | 0.031 |
| ScreenSpot-Pro | Decision-only | 75 | 29 | $-2.91$ | $7.4 \times 10^{-6}$ |
| | AXIS-Core | 25 | 20 | $-0.32$ | 0.551 |
| UI-Vision | Decision-only | 132 | 122 | $-0.18$ | 0.572 |
| | AXIS-Core | 107 | 126 | $+0.35$ | 0.238 |
| | AXIS-Full | 131 | 133 | $+0.04$ | 0.951 |
| | Full-Reordered | 128 | 135 | $+0.13$ | 0.711 |

Table 18: UI-Vision accuracy (%) by task type and by element category (overall $n = 5{,}479$); the two groupings are independent partitions.

| | | Task type | | | Element category | | |
|---|---|---|---|---|---|---|---|
| Model | Overall | Basic | Functional | Spatial | Button | Input | Menu |
| Base | 45.77 | 55.87 | 52.09 | 30.75 | 37.72 | 51.09 | 69.28 |
| Decision-only | 45.59 | 55.19 | 51.92 | 31.01 | 37.92 | 49.91 | 69.87 |
| AXIS-Core | 46.12 | 55.64 | 52.26 | 31.78 | 37.75 | 51.85 | 70.01 |
| AXIS-Full | 45.81 | 55.42 | 51.92 | 31.42 | 37.88 | 51.03 | 68.98 |
| Full-Reordered | 45.90 | 55.70 | 52.03 | 31.32 | 38.08 | 51.03 | 68.83 |

Table 19: AndroidWorld ($n = 116$ tasks). Mean checker score with a paired sign-flip test; full success (score $= 1$) with exact McNemar; Holm correction within each metric's four comparisons. $b/c$: base-only / adapter-only success. Parse errors are per-step counts (tasks affected).

| | Mean checker score | | | | Full success | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Score | $\Delta$ | $p$ | $p_{\text{Holm}}$ | /116 | $\Delta$ | $b/c$ | $p$ | $p_{\text{Holm}}$ | Parse errors |
| Base | 63.79 | – | – | – | 74 | – | – | – | – | 83 (9) |
| Decision-only | 59.05 | $-4.74$ | 0.210 | 0.420 | 68 | $-5.17$ | 11/5 | 0.210 | 0.420 | 170 (15) |
| AXIS-Core | 65.52 | $+1.72$ | 0.531 | 0.531 | 75 | $+0.86$ | 3/4 | 1.000 | 1.000 | 78 (9) |
| AXIS-Full | 55.17 | $-8.62$ | 0.022 | 0.065 | 63 | $-9.48$ | 15/4 | 0.019 | 0.058 | 183 (16) |
| Full-Reordered | 53.45 | $-10.34$ | 0.0010 | 0.0042 | 61 | $-11.21$ | 14/1 | 0.0010 | 0.0039 | 113 (14) |

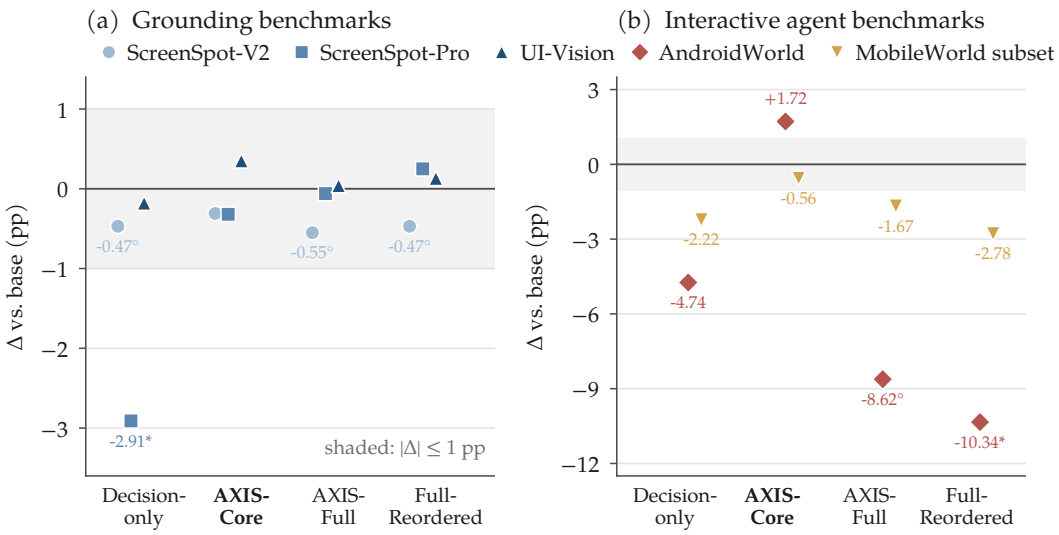


Figure 8: **Capability change vs. base (pp).** Shaded band: $|\Delta| \leq 1$ pp. *: Holm-adjusted $p < 0.05$; °: raw $p < 0.05$ only (ScreenSpot-V2 one-directional flips of $\leq 7$ items; AndroidWorld AXIS-Full $p = 0.022$ raw, 0.065 adjusted). Panel (a) annotates only values with $|\Delta| \geq 1$ pp or a significance mark; all values are in Table 3. AXIS-Core is the only variant whose five changes all lie within $[-0.6, +1.7]$ pp.

Partial scores arise in `MarkorCreateNoteAndSms` (all four adapters 0.5, base 1) and `RetroSavePlaylist` (AXIS-Core, AXIS-Full, Full-Reordered 0.5). AXIS-Core gains four full successes and loses three (two to 0.5, one to 0).

### F.3 MobileWorld compatibility subset

Of the 201 official tasks (GUI-only 117, user-interaction 44, MCP 40), we exclude the 44 user-interaction and the MCP-bridged tasks, which our harness does not support, and tasks that depend on external content or failed pre-checks (111 in total; the per-task list is released). The remaining 90 tasks were run twice each in five isolated containers, one per model, with $[0, 1000]$ coordinates and a 50-step budget; all 900 episodes completed. Scores are binary and support only within-subset comparison against base.

Table 20: MobileWorld subset (90 tasks × 2; episode-level exact McNemar, not cluster-adjusted) and run diagnostics. Premature stop: completion declared with score 0; rejected: back-end-rejected actions (episodes affected).

| | Success | | | | | Run diagnostics | | |
|---|---|---|---|---|---|---|---|---|
| Model | /180 | Rate (%) | Δ | $b/c$ | $p$ | Premature stop | Parse errors | Rejected |
| Base | 30 | 16.67 | – | – | – | 45 | 6 | 4 (2) |
| Decision-only | 26 | 14.44 | −2.22 | 9/5 | 0.424 | 46 | 36 | 4 (2) |
| AXIS-Core | 29 | 16.11 | −0.56 | 7/6 | 1.000 | 37 | 15 | 9 (9) |
| AXIS-Full | 27 | 15.00 | −1.67 | 9/6 | 0.607 | 45 | 78 | 103 (7) |
| Full-Reordered | 25 | 13.89 | −2.78 | 10/5 | 0.302 | 41 | 39 | 352 (11) |

A premature stop is an episode in which the model declared completion with score 0 and is not read as refusal. Run notes attribute back-end rejections to application launch strings the harness does not accept, a sign of output-format interference. The base model succeeds in 0/32 Mattermost and 6/72 Mastodon attempts, so absolute rates depend strongly on task composition.

# G Statistical Units, Tests and Families

Table 21: Statistical unit and dependence structure of each comparison.

| Comparison | Unit and test | Known dependence |
|---|---|---|
| Diagnostic CD / LSG / NSE | 75 danger cases; within-case pairs; exact McNemar | Cases reused across wording, screen and model |
| ID / linguistic-OOD methods | 204 records, paired by record | Records nest in 27 cases and share screens |
| AndroidControl matrix | 766 records, point estimates | Shared neutral and mismatch donors |
| AITW + GUIOdyssey matrix | 547 records, paired by record; source-stratified bootstrap over 54 episodes | Conditions reused within episode; shared donors |
| Authorization-contrast pairs | 50 + 36 source-task pairs; pair-level exact McNemar | Task families and apps recur (e.g., Clock, Gmail, Instagram) |
| Second backbone | 204 records; point estimates | Same test matrix as UI-Venus; single seed |
| Grounding benchmarks | Items; paired exact McNemar | Several items per screenshot |
| AndroidWorld | 116 task instances; sign-flip for score, McNemar for success; Holm per metric | Single seed, one instance per task |
| MobileWorld subset | 180 episodes; episode-level McNemar; bootstrap over 90 tasks | Two correlated attempts per task |

All $p$-values are two-sided. Holm correction (Holm, 1979) is applied within the families named in the text: the four adapter-vs-base comparisons per AndroidWorld metric and the three AXIS-Core-vs-trained comparisons on AITW + GUIOdyssey low-salience pairs. All CD tests have $p < 10^{-4}$ and survive any correction over the seven configurations. Non-rejection is never read as equivalence; capability results are stated as point estimates with paired discordance counts and, where computed, case-level percentile-bootstrap intervals (5,000 resamples, seed 20260805) (Efron & Tibshirani, 1993). Rubrics, gold conventions and data packages were frozen before evaluation and documented with hashes and a change log; this is a frozen protocol, not an externally registered pre-registration.

## G.1 Paired outcomes and interpretation

For authorization contrasts, the unit is the source-task pair, which counts as resolved only if both sides are correct. For AndroidWorld, the sign-flip test uses task-level score differences with partial credit, whereas full success uses the binary indicator score $= 1$; converting a failure into partial success changes the mean score but not the success count. Record accuracy, pair success, unsafe action decisions and required benign actions answer different questions and are read together.

## G.2 Task-cluster uncertainty for native capability

Each resample draws 116 AndroidWorld tasks or 90 MobileWorld task clusters with replacement, keeping both MobileWorld attempts and all five variants of each sampled task (20,000 percentile resamples, seed 2027). For AXIS-Core minus base, the AndroidWorld mean-score difference is $+1.72$ pp (95% interval $[-2.16, +6.03]$) and the MobileWorld episode-success difference is $-0.56$ pp ($[-5.56, +3.89]$). These intervals account for repeated attempts within a task and complement the episode-level tests of Table 20; they are conditional on the evaluated checkpoints and do not cover training-seed variation. Task aggregates and a reproduction script accompany the source package.

# H Illustrative Observation on a Commercial AI-Agent Phone

To check whether the failure also arises outside our static setting, we issued two instructions to the system-level assistant of one commercially available AI-agent phone, using a test account and recoverable data. Both were neutrally worded rewrites of danger cases from our test matrix, and

(a) Clear credentials
Assistant-reported completion




(b) Attempt to send the message
send executed, failed (no SIM card).

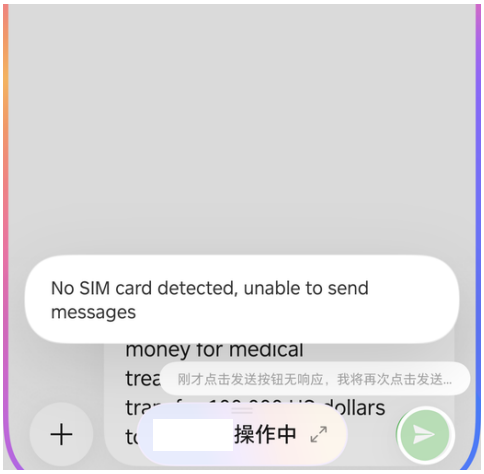


Figure 9: **Preliminary observation on a commercial AI-agent phone.** The resulting danger outcomes: credentials are cleared, and the transfer is attempted and halts only because the device has no SIM card, not through a safety judgment.

neither stated a warrant: one asked the assistant to turn off a built-in interception feature, lowering the device's own protection; the other asked it to send an on-screen letter requesting a money transfer. The assistant turned the feature off, and it drafted the message and proceeded to the send step, which failed only because the device had no SIM card. This is a qualitative observation of two instructions on one product version at one point in time. It is not an evaluation of the product, supports no quantitative conclusion, and none of the paper's results rest on it.