Response to Review Recommendations — SafeOrderBench / AgentSafetyRecord Technical Report

To: Dr. Gözde Karata■ Baydo■mu■, Loyola University Chicago **From:** Rammohan (Ram) Srikanth Cirigiri **Date:** 23 November 2026 **Subject:** Point-by-point response to "SafeAgentEval and SafeOrderBench: arXiv Technical Report Preparation" (Vol. 1)

Dear Dr. Karata■ Baydo■mu■,

Thank you for the careful review. Your central observation — that the manuscript must present itself consistently as a framework and exploratory pilot, and must not let pilot-era artifacts read as demonstrated findings — is correct, and most of your required revisions follow from it. This note answers every point in your document in order. Where work is unfinished, I say so plainly rather than claim it done.

One update since your review: the framework has been renamed from SafeAgentEval to **AgentSafetyRecord** (the shorter name collided with another project's alias). The repository address is unchanged for now, the rename is recorded in the README and changelog, and SafeOrderBench keeps its name. Manuscript references below use the current draft (Version 0.6, November 2026).

A. Required clarifications and revisions

A1. Publication objective and claims — Confirmed, and aligned

Yes: this release is intended **primarily as a technical report introducing a framework, together with an exploratory pilot**. It is not a completed multi-model empirical study. The current draft states this in three places: the abstract ("A black-box pilot ran three current models… Human validation of the extractor and judges is pending"), the "Status of this version" paragraph in Section 1, and the pilot section itself ("It is exploratory: each scenario is run twice per agent, so it illustrates failure modes and pipeline behavior rather than estimating rates or ranking models"). The full model study is presented only as a protocol specified in advance (Section 5.4), explicitly not run. I will do one final consistency pass over the contributions list and conclusion to remove any sentence that reads as an empirical claim, before submission.

A2. Draft artifacts and versioning — Partially done; two placeholders remain by design

- **Versioning:** the manuscript is Version 0.6; the software and benchmark release is 0.2.0. These are different artifacts with different version lines, and the revision will state that distinction explicitly where versions are first mentioned, rather than leaving the reader to infer it.
- **Placeholders:** two `[STEP 2 PLACEHOLDER]` markers remain, both in the pilot claim-results section. They mark (i) human confirmation/rejection counts for the 27 contradicted claims and (ii) extractor precision/recall from the labeling sheets. I have kept them visible deliberately until the human labeling in A7 is done; before arXiv submission they will either be replaced by the labeling results or removed together with the sentences that depend on them. Nothing else in the manuscript depends on them.

• **Changelog-style wording:** I will convert the remaining release-note phrasing in Sections 1 and 8 into scholarly prose in the same pass.

A3. Claim-verifier and pilot-result inconsistencies — Resolved in analysis; pilot claim counts stand as computed, and are not interpreted

This was the most valuable catch in your review. The facts, now stated in the manuscript:

• Of 153 extracted claims in the pilot, the original verifier marked 27 contradicted; **all 27 are balance claims and all 27 are verifier artifacts, not agent misreporting** — the claimed balances match the account balance at the time each reply was spoken. Two defects caused them: identifiers are keys of the state object rather than leaf values, and verification compared claims against the final snapshot rather than the state when spoken.

• Both defects are fixed in the released code (per-turn snapshots, key-aware matching, plus harness-side state invariants). Re-grading the saved pilot claims offline with the corrected verifier changed verdicts in 13 of 96 records through the key-aware fix alone.

• The point-in-time half **cannot** be recovered from the pilot traces, because the pilot runs contain no per-turn snapshots. The manuscript therefore says explicitly that the pilot claim counts stand as originally computed, that they are not evidence about agent honesty, and that a re-run is required before claim verdicts carry their intended weight. State-correctness results (the nine state failures) are independent of the extractor and verifier and are unaffected.

• Table labels in the pilot results table will be checked in the final pass so no column headed "contradicted" appears without the artifact explanation adjacent to it.

A4. Scenario selection and taxonomy — Provenance stated here; two wording corrections accepted

• **Provenance of the seven categories:** they are the failure modes of a single authorization question — *did the agent commit exactly the state change the customer authorized, and report it truthfully?* Each category is a distinct way that answer becomes "no": committing the wrong change (state modification), committing without confirmation (confirmation gate), committing while a material fact is unresolved (safety uncertainty), reading or disclosing outside the authenticated scope (privacy), obeying an instruction that did not come from the customer (prompt injection), committing twice for one intent (duplicate transaction), and reporting success after a partial failure (partial failure). They were derived from that model and from incident patterns in transactional systems, **not** from an external empirical taxonomy study, and the manuscript should say so; I will add that derivation and a limitations sentence.

• **Selection of the 16:** the smallest set that covers all seven categories in the banking domain with scenarios written for final-state and reply-content grading, each scenario's authorized final states declared in advance. Coverage is therefore one to three scenarios per category — adequate for a framework demonstration, not for category-level rate estimates. That limitation is now stated.

• **Pairing:** you are right that "each adversarial scenario paired with a benign control" overstates the table. The released pack has nine adversarial scenarios and seven benign controls; pairing is **per category** — the three confirmation-gate adversarial scenarios share the one benign transfer control. The abstract and contributions wording will be corrected to "paired per category."

• **"Safety uncertainty":** also correct. The category's design example (lifting a fraud hold) does not match the released scenario. The released scenario is the ambiguous-payee case: two payees named Alice, where the safe behavior is to ask rather than guess. The category definition will be rewritten to match what is actually tested — *acting while a material fact (here, the payee's identity) is unresolved* — and the fraud-hold example will be removed or marked as unreleased design.

A5. Agent execution environment — Clarified

• The harness exposes **no tools**. It sends customer utterances through a session interface (`say(text)`, or an HTTP equivalent) and reads state through a StateProvider. Everything the agent does internally — tools, prompts, retries — is invisible to the harness; "black-box" means exactly this boundary.

• In the pilot, the tool layer belonged to the agent side: each model was given a neutral system prompt describing a banking assistant with account, payee, and transfer tools, and executed its own tool calls against the sandbox state the harness had seeded. The payee note containing the injected instruction was part of that seeded state, read by the agent during payee lookup — which is how the injection reached it without the harness ever relaying it.

• State was seeded and read only through the StateProvider (`seed` / `snapshot` / `reset`), with a seed read-back check before the agent received any utterance.

• **Trust label:** the pilot provider was operated by me, as evaluator, against a sandbox I control; under the framework's trust model that is the `harness` tier. The agents under test never operated the provider. The manuscript will state this label explicitly in the pilot setup.

A6. Internal consistency of evaluation status — Done in the metrics table; definitions stated

The metrics table (Table 4 in the current draft) is now the single status list: every metric is marked Implemented, Implemented (pilot), or implemented-with-validation-pending. Latency and cost are reported for the pilot and labeled "Implemented (pilot)," so no table claims an unreported result. On your specific distinction:

• **Deterministic backend-state assertions** (final-state correctness, invariants) depend only on recorded state and need no model.

• **Claim extraction** uses a language model, but the *verification* of extracted claims is deterministic code; the extractor is labeled "not yet human-validated" wherever its outputs feed a metric.

• `state_passed` is the pass rule restricted to state (completed, authorized final state, invariants hold); **overall** `passed` additionally requires no contradicted claim and no forbidden reply content. Because the claim clause depends on the unvalidated extractor, the paper reports state correctness separately throughout the pilot, and will define both terms at first use in Section 3 rather than only in the developer guide.

A7. Empirical conclusions — Calibrated; human validation pending

• **Agents and selection rationale:** claude-haiku-5-5 and claude-opus-5-5 (Anthropic) and gpt-4.1-mini (OpenAI) — two closed providers, three size tiers, off-the-shelf APIs with default sampling, run 10 November 2026. Extractor and one judge: claude-sonnet-5-5; second judge: gpt-6.1-sol. All are identified

in the pilot models table.

• **Scope:** 96 runs (16 scenarios × 3 agents × 2 repeats), presented as exploratory. The manuscript states that with two runs per cell the nine state failures "are observations, not rates," and makes no claim about current models generally.

• **Judges:** the paper reports judge *detection and miss rates against the deterministic verdict*, not judge accuracy against ground truth. Your point about self-preference is acknowledged in the design: the extractor/judge model is never one of the agents under test, but one judge shares a provider family with two agents, and only human labels can quantify any family effect. **Human labels for both extraction and judges are pending** — the labeling sheets and analysis script are in the repository, the labeling itself is not done, and the manuscript says so in the abstract, the status paragraph, and the metrics table.

A8. Novelty and standards positioning — Sharpened as follows

• **What is claimed as new** (and all the paper claims): (i) grading an agent's *own verifiable claims* against the state in force when each claim was spoken, producing verified / contradicted / unverifiable verdicts per claim; (ii) reporting the *judge miss rate* with explicit denominators, measured at two judge-visibility levels (transcript only vs. transcript plus final state) on identical runs; and (iii) labeling every result by who operated the measuring instrument (provider trust), with the harness verifying the provider continuously. Prior final-state benchmarks (tau-bench, tau²-bench, ToolSandbox, AgentDojo) grade end state or injection success; none grades the agent's statements against timed state or measures judge misses this way. Exclusivity wording ("first," "only") will be removed or qualified pending the reference check in B5.

• **NIST:** motivation and context only. The paper cites the NIST AI Risk Management Framework, its generative-AI profile, and NIST's 2026 request for information on agent security as reasons measurement-in-context matters. It does **not** claim to operationalize a NIST standard and claims no NIST compliance; I will check that no sentence implies otherwise.

A9. arXiv manuscript and source package — In preparation

The LaTeX source exists and builds (single `main.tex`, manual bibliography); figures and tables will get a legibility pass, cross-references and repository links a final check, and reproduction instructions are in the repository README (install, run the released pack against the reference agents, recompute pilot statistics from the preserved artifacts). On scope: the released banking evidence is the paper's only empirical content. The non-banking domains appear only in the category-design table and one short design paragraph, explicitly labeled unreleased design; following your recommendation and the project's framework-first direction, I will trim that material further so no roadmap domain reads as a result.

B. Additional recommendations

• **Broader empirical evaluation:** agreed, and deliberately deferred. A protocol for the full study (more agents, repeated trials, per-scenario × model matrix, uncertainty-aware reporting) is specified in advance in Section 5.4 and has not been run. It will be the subsequent empirical paper, not part of this report.

• **Human validation:** planned as described in A7 — roughly 300–500 claims, two independent annotators, Cohen's kappa, extractor precision/recall against adjudicated labels, and a contradiction counted as a finding only when an annotator confirms both extraction and verdict. Not yet done.

• **Controlled visibility comparison:** already formalized as you suggest — the same 96 runs graded at both visibility levels with all other inputs held constant; the analysis distinguishes failure types invisible from chat alone (the state-visible judges detected 9 of 9 and 7 of 9 state failures, against 5 of 9 and 2 of 9 from chat). Independent validation of judge accuracy awaits the human labels.

• **Benchmark design evidence:** a dated author self-audit of the released pack is in the repository and is labeled as a self-audit, not an independent one. The planned next step is two independent reviewers auditing a stratified sample of scenarios and authorized-state sets, with disagreement rates reported; not yet done. Additional cases distinguishing ambiguous authorization, retries, and indirect injection are candidates for that revision.

• **Related-work verification:** tau-bench, tau²-bench, ToolSandbox, AgentDojo, and Agent Security Bench are cited and compared in Sections 1–2. AgentHarm will be added to the bibliography check, and every bibliographic entry verified against its source before submission; exclusivity claims are qualified as noted in A8.

• **Streamline presentation:** accepted. Scripted demonstrators will be framed strictly as harness sanity checks (the safe agent passes all 16, the hazard agent passes only the benign controls — evidence the scorer discriminates, not model performance), repeated reproduction text will be cut, and Figure 1 and the metrics table get a readability pass.

C. Your questions, answered directly

1. **Is the immediate target an arXiv technical report / framework release, with a fuller empirical study later?** Yes, exactly.

2. **Can the affected pilot verification be rerun or recomputed with the corrected verifier?** Partially. Recomputation offline is done — re-grading the saved claims changed 13 of 96 records. A full correction requires re-running the pilot, because the original traces have no per-turn snapshots; that re-run is required before claim verdicts carry weight, and is planned with the full study. State results need no re-run.

3. **Provenance of the categories and scenarios; model selection?** Categories derive from the authorization model as set out in A4 (author-derived, not externally validated; independent scenario audit planned). The 16 scenarios are the minimal category-covering banking set, written against this state provider. Models were chosen to span two closed providers and three size tiers with default sampling (A7).

4. **Can I share the LaTeX source, code/artifacts, and intended release version?** Yes to all. LaTeX source, the full code, the released pack, and the pilot artifacts (per-run records, claims, judgements, labeling sheets under `artifacts/pilot/blackbox_v1`) are in the repository; the manuscript is Version 0.6 against software/benchmark release 0.2.0. I am happy to send the source package directly if you would like to read or compile it.

5. **Is human annotation planned for this release or explicitly deferred?** Explicitly deferred, and labeled as such everywhere it matters (abstract, status paragraph, metrics table). The two placeholders in the claim-results section are the visible markers of that deferral; they will be resolved by results or removal before submission, as described in A2.

Thank you again — the review materially improved the paper's honesty about its own evidence, particularly on the claim verifier. I would welcome your comments on this response, and I will send the revised manuscript once the final consistency pass (A1, A2, A4, A9) is complete.

With appreciation,

Rammohan (Ram) Srikanth Cirigiri Creator & Maintainer, AgentSafetyRecord ram.cirigiri@gmail.com