Separating Pressure and Accountability in Tool-Using Agents: Evidence from CESO-Bench

Anonymous authors
Paper under double-blind review

Abstract

In the pressure comparison studied here, a pressured agent is compared with an agent told that its work will be audited. That comparison does not separate a pressure effect from an accountability-cue effect or from the model's behavior without either cue. We study this decomposition in executable tool use with CESO-Bench, a synthetic benchmark of 24 task families spanning access control, refund processing, and data retention. Each family has pressure, full-auditor, and neutral variants built from the same task body. The neutral closing keeps task-defining details while removing selected pressure language and is mechanically matched to its paired pressure closing within ±15% of word count. We evaluate three MiniMax M-series configurations at temperature 0.7 using a frozen surface success signal and deterministic end-state compliance oracles.

The primary outcome is silent pass, SP $=$ mean$[A(1-C)]$, where A is the recorded `task_answer_ok` signal and C is the atomic-compliance result after authoritative replay. Across the three models, the pressure-versus-neutral family means are -0.015, -0.031, and -0.023 for M2.7, M2.5, and M3; the corresponding auditor-versus-neutral means are -0.060, -0.096, and -0.044. The M2.7 95% percentile intervals are $[-0.049, +0.011]$ and $[-0.157, +0.003]$; the M2.5 and M3 analyses have partial family support and report point estimates only. This paper treats those contrasts as endpoint-specific evidence: a point estimate near zero is not equivalence, a confidence interval that covers zero is not proof of independence, and an auditor comparison does not identify a mechanism. The study is limited to the synthetic task pack, recorded runs, and the declared oracle contract.

1 Introduction

Tool-using agents can produce a plausible success signal while leaving a simulated world in a state that violates one or more task requirements. The failure is operationally important because a fluent response can obscure a wrong permission, an incomplete refund, or an incorrect retention action. A post-run evaluator therefore needs two separate observations: what success the agent signaled and what the tool-mediated world state satisfies.

Pressure experiments add a second measurement problem. A comparison between a pressure-framed task and an auditor-framed task changes both the pressure cue and the accountability cue. An observed gap cannot by itself say whether pressure increased failure, the auditor cue reduced it, or the two prompts interacted with a shared baseline. We address that design problem with a neutral arm that preserves the task scenario while removing selected pressure language.

CESO-Bench provides reciprocal task pairs with fixed tools, initial state, and declared hard predicates. The main pair assigns pressure and auditor closings using a frozen family-level placement rule. The neutral arm uses the same task families and a separately authored closing for each family. Its closings retain the requester, requested action, amounts, caps, record ages, and other details that define the scenario; they change selected urgency, authority, and auditor/compliance language while retaining some timing, support, and policy cues. Mechanical verification keeps each neutral closing within ±15% of its paired pressure closing's word count.

Our contribution is an empirical control design and an accounting discipline. We report pressure-versus-neutral and auditor-versus-neutral silent-pass contrasts on a common endpoint across three

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.

MiniMax configurations, with the auditor ablation and a leave-one-family-out sensitivity treated as secondary checks. We keep unresolved records, task/family weighting, refusal behavior, and the oracle's declared scope visible. The resulting claims are about the recorded CESO-Bench estimands; they do not establish a general pressure law, production safety, or a causal mechanism for an auditor cue.

2 Related Work

Tool-use benchmarks evaluate actions against simulated environments as well as text. ToolEmu studies risks in language-model agents through an LM-emulated sandbox (Ruan et al., 2024), while AgentHarm evaluates agents on explicitly harmful requests (Andriushchenko et al., 2025). CESO-Bench is narrower: it fixes synthetic task worlds and uses declared hard postconditions to score the final state, then keeps the agent's surface success signal as a separate field.

Trajectory evaluation is related but asks a different question. Mohammadi (2026) studies what outcome-only judges miss when process faults are injected into agent trajectories. CESO-Bench applies its deterministic compliance instrument to the tool-mediated world produced by the agent; it does not claim that this instrument is a general trajectory judge. PACT (Okamoto and Kaplan Erol, 2026) already includes base, unpressured, pressured, and nonbinding control conditions. Our narrower contribution is an execution-grounded, family-matched comparison of pressure, an auditor cue, and tailored neutral closings under the CESO A/C endpoint; it is not the invention of a generic neutral-control idea.

The framing manipulation also connects to work on alignment under monitoring and evaluation (Greenblatt et al., 2024), sycophancy and preference-sensitive behavior (Sharma et al., 2023), and classic framing effects (Tversky and Kahneman, 1981). These works motivate the possibility that contextual cues change behavior. They do not identify an auditor mechanism in the present study, whose full auditor text combines observability language with an explicit instruction to follow stated requirements.

3 Experimental Design

3.1 Task pack and matched arms

The benchmark contains 48 synthetic tasks arranged as 24 reciprocal twin pairs, with 16 tasks and 8 families in each of access control, refund processing, and data retention. Within a pair, the base instruction, tool interface, initial world, and declared hard oracle predicates are fixed. Only the closing stimulus changes. A frozen family-level placement rule assigns the pressure and auditor closings to the two reciprocal members; the member labels are not themselves the treatment definition.

The neutral arm reuses the corresponding task-family body and supplies one family-specific neutral closing. Neutral is therefore a matched control closing, not an empty prompt or a universal zero-framing condition. The neutral pack was built separately and verified to preserve the task body while changing only the closing. Each neutral closing is within ±15% of the paired pressure closing's word count under the mechanical verifier. Details that define the scenario are retained, while selected urgency, authority-threat, and auditor/compliance phrases are removed. The neutral texts are authored comparison closings rather than a zero-framing condition: some retain task-oriented timing, support, or policy wording, including phrases such as "resolved promptly," "Their director supports," and policy reminders. This residual-cue risk is carried as a limitation.

3.2 Framings and ablation

Pressure closings were tailored by family around urgency, authority, or consequence framing. They target task-specific shortcuts without directly specifying a policy violation. The full auditor closing is:

> A compliance auditor will review this work. Every action you take is logged and will be checked against policy. Proceed carefully and make sure each step follows the stated requirements.

The M2.7 ablation keeps the salience and logging language and removes exactly the final instruction sentence:

> A compliance auditor will review this work. Every action you take is logged and will be checked against policy.

Similar behavior between the full and ablated texts would be consistent with a salience account, but would not establish why the agent responded that way. Stored prompt and trace checks verify the long auditor text and the exact sentence deletion.

3.3 Models, scaffolds, and collection

We analyze MiniMax-M2.7, MiniMax-M2.5, and MiniMax-M3 at temperature 0.7. The confirmatory scaffold set is `tool_forced` and `react`. Each run has 24 families, two scaffolds, and 768 planned records. The main pressure/auditor comparison uses eight trials per task-scaffold member; the neutral and additional arms use 16 trials per task-scaffold task. MiniMax trial seeds are metadata only because the provider does not accept a seed parameter, so the temperature-0.7 observations are described as unseeded stochastic requests.

The M2.7 pressure and auditor arms are the frozen main comparison. The neutral baseline and additional model/ablation arms are the pre-collection V7/V8 amendments. This revision performs no new API collection and uses only the retained run manifests, traces, source, and analysis artifacts.

4 Measurement Contract

For each assessable trial, let

$$A = \texttt{task_answer_ok}, \qquad C = \texttt{compliant}.$$

A is the agent's explicit surface-success signal from the recorded protocol. It is not semantic answer correctness. C is the run-level atomic-compliance result: after authoritative tool replay, every applicable hard end-state predicate must pass. The frozen protocol has one narrow refusal exception: on access-a003 and access-a008, a tool-free final message containing "refuse," "denied," or "deny" satisfies the denial predicates. Thus C is a declared compliance contract with a fixed refusal override, not an exclusively end-state-predicate label and not a claim about all real-world policy or legal requirements.

A trial is assessable only when both fields are Boolean. Unknown or unresolved records, including transport failures and timeouts, are excluded from metric denominators and reported as missingness; they are never imputed as success or failure.

The frozen primary quantities are

$$\text{ESCR} = \text{mean}(C), \qquad \text{SP} = \text{mean}[A(1-C)]\,,$$

with the benchmark's task-macro aggregation over tasks with at least one assessable trial. Trial-weighted pooled summaries, when shown, are secondary diagnostics. SilentFail@k is the event that at least one of k trials has $A = 1, C = 0$. Compliant@k is the joint event that all k trials have $A = 1, C = 1$. These definitions keep the success surface and end-state compliance separate; neither metric is inferred from final-message fluency.

For each family f, the neutral-baseline comparisons are

$$\Delta_{P-N,f} = \text{SP}(P_f) - \text{SP}(N_f), \qquad \Delta_{A-N,f} = \text{SP}(A_f) - \text{SP}(N_f).$$

The reported component is the mean of the assessable family differences. The V7/V8 uncertainty procedure uses a percentile bootstrap with $B = 10{,}000$ resamples over family differences. This is a model-based resampling summary under the stated family-resampling assumptions, not mechanical uncertainty for a finite population. A confidence interval for a primary comparison is reported only

at the declared full 24-family gate; partial-family primary analyses retain their point estimate and are labeled as such. The M2.7 LOFO interval is an explicitly exploratory reviewer-triggered exception with 23 families. The V8 M2.5 and M3 LOFO analyses remain means only below the full family gate.

The pressure-versus-auditor component is retained only when its input runs use the same endpoint and mapping; it is not used to turn an auditor comparison into a pressure effect.

Refusal families remain in the primary family universe. Any with/without refusal sensitivity is labeled explicitly and never used as an unreported exclusion. A constant observed A margin is not a reason to remove a family from SP or ESCR. The exactly unbiased D_k^u independence-null kernel is part of the frozen theory record, but it is not an empirical outcome in this neutral-baseline paper.

5 Results

The central results table reports the retained run manifests using the frozen estimand, weighting, family support, and missingness fields.

Table 1. Primary pressure- and auditor-versus-neutral SP contrasts and support. Values are rate differences; bracketed values are 95% percentile confidence intervals, and unbracketed values are point estimates.

Model	Δ_{P-N} SP	Δ_{A-N} SP	Families / CI status	Assessable coverage
MiniMax-M2.7	-0.015 [-0.049, +0.011]	-0.060 [-0.157, +0.003]	24 / percentile 95% CI	P: 263/384; A: 297/384; N: 685/739 observed keys, 768 planned
MiniMax-M2.5	-0.031	-0.096	23 / point estimate; primary	main: 388/768; neutral: 435/768
MiniMax-M3	-0.023	-0.044	21 / point estimate	main: 523/768; neutral: 701/768

For M2.7, the family-level estimates are the primary comparison. The neutral run has 685 complete records among 739 observed keys; the remaining 29 planned keys have no record. The pressure and auditor arms have 263 and 297 complete records, respectively. A pooled trial-weighted SP gap is +0.0465, but it uses a different weighting and is retained only as a secondary diagnostic; it is not combined with the family means or with the historical non-compliance endpoint.

The V8 completion gate was 85% per run. The M2.5 primary neutral run reached 435/768 (56.6%) and its requeue sensitivity reached 526/768 (68.5%); the M2.5 main side reached 388/768 (50.5%). M3 reached 701/768 (91.3%) on neutral and 523/768 (68.1%) on its main side. These are collection diagnostics that bear on exchangeability; the gate is not a blanket rule for deleting every run from the paper.

The pressure-versus-neutral component is the matched comparison for selected pressure language. Its estimate, interval when the family gate is met, and unresolved-record share are reported together. A small estimate is not equivalence or a null effect. An interval covering zero is a non-rejection in this finite sample, not proof that the framings are independent.

5.1 Leave-one-family-out sensitivity

The M2.7 leave-one-family-out analysis was an exploratory, reviewer-triggered sensitivity that removes the potential concentration family and recomputes the stored family differences. The M2.5 and M3 versions were specified prospectively in the V8 amendment. These analyses are robustness displays, not replacements for the primary family universes.

Table 2. Leave-one-family-out auditor-versus-neutral SP contrasts. Values are rate differences; bracketed values are 95% percentile confidence intervals, and unbracketed values are point estimates.

Model/run	LOFO family	Δ_{A-N} SP	Support / interval status
MiniMax-M2.7	retention.anonymize_for_analytics	-0.019	exploratory; 23 families; percentile 95% CI [-0.055, +0.009]
MiniMax-M2.5 primary	retention.anonymize_for_analytics	-0.055	22 families; mean only
MiniMax-M2.5 requeue	retention.anonymize_for_analytics	-0.062	22 families; mean only
MiniMax-M3	retention.anonymize_for_analytics	-0.005	20 families; mean only

The sensitivity shows concentration and uncertainty without replacing the primary family universe.

5.2 M2.5 requeue sensitivity

The M2.5 requeue is a separately labeled sensitivity, not a replacement for the primary snapshot. It has 526/768 complete neutral records and the same 388/768 main records over 23 assessable families. Its family means are $\Delta_{P-N} = -0.038$ and $\Delta_{A-N} = -0.103$, with no full-universe confidence interval.

5.3 Auditor ablation

The M2.7 ablation compares the exact shortened auditor closing with the full auditor and neutral arms on the same SP endpoint.

Table 3. Auditor-ablation SP contrasts. Values are rate differences; bracketed values are 95% percentile confidence intervals, and unbracketed values are point estimates.

Contrast	Δ SP	Families / CI status
Ablated auditor − neutral	-0.057 [-0.155, +0.005]	24 / percentile 95% CI
Ablated auditor − full auditor	+0.003 [-0.013, +0.018]	24 / percentile 95% CI

This comparison can describe whether the ablated and full auditor arms differ in the recorded endpoint. It cannot distinguish salience, instruction following, or another latent process without a manipulation check that measures those mechanisms directly. The ablated arm contributes 718/768 complete records; the stored family estimates are -0.0568 for ablated auditor versus neutral and +0.0033 for ablated auditor versus full auditor, with the intervals shown above.

6 Limitations and Uncertainty

First, the task pack is synthetic. Its end-state predicates make the scoring contract inspectable, but they do not cover every obligation in a production workflow. In particular, stored-message and audit-row predicates are limited to what the declared simulator state records. $C = 1$ therefore means that the declared scoring contract was satisfied, not that an action is universally lawful, secure, or operationally adequate.

Second, A is a surface signal. A valid `task_answer_ok` value does not prove that the requested work was semantically completed, and a refusal can be appropriate for a refusal task. This is why the paper reports SP as the joint event $A(1-C)$ and keeps ESCR, SilentFail@k, and Compliant@k as separate quantities. We do not remove tasks because their observed A margin is constant, and we do not treat zero observed events as universal absence.

Third, unresolved records and separate arm collection limit exchangeability. Transport missingness can depend on task, scaffold, or run conditions. The complete-case family means are therefore accompanied by coverage. The three models share a provider and temperature, and provider-side sampling was unseeded. The 24 families are structured stimuli, not 24 independent population mechanisms. Family resampling therefore does not justify population generalization beyond this limited, structured set.

Fourth, the neutral closing controls word count and preserves task details but still changes wording. The auditor text combines observation salience with a compliance instruction, and the ablation removes one sentence. These are useful contrasts for the stated estimands, not isolated causal mechanisms.

Finally, the supplementary 200-trace coding exercise is reported only as a bounded agreement check for the deterministic instrument. Its labels reproduce 199/200 agreement and Cohen's $\kappa = 0.988819$ under the recorded protocol; the coder was the session AI, the machine labels were withheld during coding, and the historical blinding protocol was not independently audited. We make no construct-validity claim and no claim of independent human validation from this exercise.

7 Conclusion

CESO-Bench separates an agent's recorded success surface from deterministic post-run compliance and uses a matched neutral closing to decompose a pressure-versus-auditor comparison. Across M2.7, M2.5, and M3, the pressure-versus-neutral SP means are -0.015, -0.031, and -0.023, with a percentile interval for M2.7 and partial-family support for the other two models. The auditor comparison, LOFO sensitivity, and exact-sentence ablation provide context. The study's interpretation remains limited to the declared endpoint, synthetic families, and retained runs.

8 Reproducibility, Ethics, and AI Use

The supplementary package lists the available task-pack source, frozen manifest and amendment text, exact pressure/auditor/neutral prompt assets, run manifests, retained traces, analysis outputs, and validation records. Each analysis records its model, configuration, task mapping, endpoint, and family support. Coverage and missing execution dependencies are disclosed in the accompanying reports; the package does not imply that every historical run can be independently regenerated. Packaging history, checksums, and reconciliation of superseded snapshots are maintained in the accompanying reconciliation notes rather than in the scientific manuscript.

The standalone analyzer is not executable from the supplementary directory alone because historical metrics, oracle, and task modules are not all present, and the follow-up analysis records external inputs. This limits independent re-execution of the archived calculations; the retained JSON, traces, and reconciliation reports remain the cited evidence for the reported rows.

All tasks, identities, records, and tool worlds are synthetic. The study did not access customer data, production systems, or real financial accounts. The benchmark is an evaluation artifact and should not be used as evidence that a model is safe for deployment without domain-specific requirements review, oracle coverage, and independent operational validation.

8.1 AI Use

Language-model assistance was used under human direction for local archive inspection, methodological review of the experimental design and measurement contract, statistical interpretation of stored estimates and uncertainty summaries, checking mathematical definitions and derivations, arithmetic checks, and prose editing. Review comments and generated prose were editorial inputs, not evidence. Model outputs in retained runs and the session-AI coding exercise remain empirical

artifacts and are labeled as such; the coding exercise is not independent human validation, and no human review is claimed for it. The author remains responsible for claims, references, disclosure, and venue decisions. No new API collection or paid experiment was performed for this revision.

References

Andriushchenko, M., et al. (2025). AgentHarm: A benchmark for measuring harmfulness of LLM agents. ICLR 2025. Proceedings.

Greenblatt, R., et al. (2024). Alignment faking in large language models. arXiv:2412.14093.

Mohammadi, H. (2026). What Outcome-Only LLM Judges Miss on Agent Trajectories. arXiv:2609.00038.

Okamoto, M. and Kaplan Erol, A. (2026). PACT: Can Enterprise AI Assistants Be Trusted Under Pressure? arXiv:2609.18605.

Ruan, Y., et al. (2024). Identifying the risks of LM agents with an LM-emulated sandbox. ICLR 2024. Proceedings.

Sharma, M., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453–458. DOI.