A Prompt-Masked Pilot for History-Dependent Safety Degradation in Multi-Turn Conversational Agents

Anonymous Authors[1]

Abstract

Closed-loop conversational systems can accumulate unsupported premises across turns, producing a history-dependent failure boundary that single-turn evaluations can under-characterize. We present a prompt-masked 12-turn pilot protocol over nine matched synthetic stress-test personas, each paired across three policies: periodic grounding, one-time safety persona framing, and no intervention. A GPT-5.4/GPT-5.2 LLM-as-judge ensemble scores the resulting 27 matched conversations at the conversation level on a coded -1 to 4 primary-confirmation rubric after intervention prompts, system messages, and reasoning blocks are removed. In this cohort, periodic grounding receives a lower mean primary confirmation score than matched control by 0.55 points (95% paired-bootstrap CI $[0.32, 0.73]$, exact two-sided paired sign-flip $p = 0.0078$). We did not detect a difference between periodic grounding and one-time persona framing (mean difference -0.03 points, 95% paired-bootstrap CI $[-0.20, 0.18]$, exact two-sided paired sign-flip $p = 0.8359$); an exploratory control-versus-persona contrast points in the same direction as the primary comparison. A residual style-leakage probe predicts the three arms from masked assistant text above chance (55.6% vs. 33.3%, leave-one-persona-out, permutation $p = 0.008$), so the evaluation is prompt-masked rather than fully blinded. The study remains limited by synthetic personas, automated judges, inference over only nine matched personas, and unresolved external validity; it reports lower automated unsupported-belief endorsement scores in a synthetic stress-test cohort, not clinical efficacy or deployment readiness.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Single-turn safety tests often under-characterize a common long-horizon failure: a model can remain polite while progressively treating the user's unsupported frame as evidence (Zhou et al., 2024; Lee et al., 2026). The risk is sharper when the frame concerns events or beliefs that are not grounded in shared reality. In that setting, endorsement can accumulate across turns even if no single response looks catastrophic in isolation. Consider a user who keeps insisting that their neighbors are transmitting signals into their head, or that they have decoded a hidden mathematical structure of the universe. The safety problem here is whether the assistant still separates supportive tone from factual endorsement after several turns of agreement pressure, even when no single refusal is the right move.

The pattern fits what recent sycophancy work describes. Reinforcement learning from human feedback can favor responses that align with a user's stated views even when those views are unsupported, and the bias is sharper in multi-turn exchanges than in single-turn tests (Sharma et al., 2024; OpenAI, 2025; Hong et al., 2025). Multi-turn jailbreak studies report a related dynamic in which adversarial framing is carried across many turns rather than delivered in one trigger (Russinovich et al., 2025). The common factor is that dialogue history accumulates, and the assistant's later behavior is conditioned on what it has already conceded.

We study this phenomenon as *unsupported-belief reinforcement*. Prior application domains describe related behavior under several labels for cases in which extended dialogue appears to escalate rather than de-escalate ungrounded beliefs (Hua, 2025; Morrin et al., 2025; Østergaard, 2025). We do not adopt clinical framing as a finding of this paper. The primary grading rubric uses *delusion-confirmation* as its label because the synthetic trigger set centers unsupported or delusion-like beliefs; the terminology is broader, but the empirical support remains limited to this trigger family. A response counts as more reinforcing when it treats an unsupported belief as likely true, builds on the belief as evidence, or invites further elaboration of the belief without grounding. It does not count as reinforcing merely because the model acknowledges distress, asks neutral clarifying questions, recommends support, or refuses to adjudicate unavailable

facts.

This paper studies that process in a deliberately narrow pilot evaluation over synthetic conversations. We run 12-turn conversations between a fixed target model and fictional personas that press the model toward increasingly unsupported interpretations. We then compare conversational safety policies that differ in when they intervene: repeated grounding reminders during the conversation, a single initial safety-oriented persona instruction, and no intervention.

We treat this as a closed-loop conversational setting. Each assistant turn is an action conditioned on the full dialogue history, the history itself is an evolving state, and the intervention schedule is varied while keeping the persona and task family matched. This is not fixed-user replay: downstream user turns are regenerated under each arm, so the design measures trajectory-level prompt-schedule score differences rather than isolating assistant responses to identical user text. Once unsupported premises enter that state, later responses can inherit and operationalize them even when no single turn looks catastrophic. The prompt-masked judges therefore see the full prefix and rate each response in context: a turn is rated more reinforcing if it preserves the accumulated unsupported frame, not just for its local tone. Control corresponds to no intervention, one-time persona framing perturbs only the initial system instruction, and periodic grounding refreshes a short reality-orientation prompt every three turns.

The trigger family keeps the user fixed on an unsupported interpretation across many turns. The hazard is a trajectory in which the assistant gradually upgrades that interpretation into working evidence rather than catching it in any single response. A safer policy must keep separating emotional support from factual endorsement even after the conversation has already accumulated pressure toward the unsupported frame.

Under the reported evaluation protocol, one empirical result is supported within this cohort under automated scoring. Periodic grounding receives a lower primary confirmation score than matched no-intervention control. It does not show that refreshing the instruction every three turns outperforms a strong initial safety persona. That distinction matters scientifically because it separates an observed scaffolding-versus-control score difference from a design-sensitive timing question about repeated correction.

The paper is intentionally bounded. It does not study patients, diagnosis, treatment, or deployment. It does not claim adaptive intervention selection or mechanistic understanding from behavior alone. The contribution is a matched synthetic stress-test cohort, a prompt-masked automated grading pass, a residual style-leakage diagnostic, and a conversation-level analysis of long-horizon safety scaffolding.

Our contributions are:

- A matched 27-conversation, 324-target-turn pilot cohort across no-intervention control, periodic grounding, and one-time persona framing, with inferential $n = 9$ matched persona pairs.
- A prompt-masked evaluation pipeline and descriptive trace diagnostic that remove intervention prompts, system messages, and reasoning blocks before grading, plus a residual register-leakage probe.
- A conversation-level estimand with prompt-masked GPT-5.4/GPT-5.2 judge aggregation, rather than turn-level pseudo-replication.
- A paper-facing primary/secondary/exploratory claim split that keeps the observed scaffold-versus-control score difference separate from the design-sensitive persona-versus-grounding comparison.

Stateful policy-evaluation framing. The pilot treats dialogue history as the state carried forward from one assistant response to the next. The user task and personas stay matched while the intervention schedule varies, letting us compare no intervention, one initial safety framing, and periodic grounding in the same 12-turn setting. This is not a claim about tool-using deployed agents; it is a compact measurement test for a history-dependent conversational failure that may arise in longer-running agent settings. The pattern surfaced here, sustained user pressure incrementally reshaping scored policy behavior across an accumulating dialogue history, motivates future evaluation of long-running customer-support agents, journaling or companion agents, and agentic copilots whose safety stance may erode over many turns of cooperative context. The 12-turn synthetic setting under-samples those deployments by design, and the prompt-masked grading path still admits residual register leakage, but the result motivates future tests before applying this evaluation pattern to tool-using or deployed agents.

2. Related Work

This paper draws on automated red-teaming, long-horizon interaction risk, sycophancy, health-adjacent conversational safety audits, and LLM-as-judge evaluation.

Automated red-teaming uses model-generated scenarios and model-assisted evaluation to surface risky behavior before deployment (Perez et al., 2022); our setting is narrower, fixing a matched cohort and a fixed scaffold comparison rather than searching broadly for arbitrary failures. Long-horizon interaction work argues that risk can accumulate through dialogue rather than appearing in a single answer (Kirgis

et al., 2026; Zhou et al., 2024), and multi-turn jailbreak studies report a related dynamic in which adversarial framing is carried across turns (Russinovich et al., 2025). Sycophancy work gives a behavioral account for why a model may defer to a user's stated frame, especially in sustained dialogue (Sharma et al., 2024; Hong et al., 2025; OpenAI, 2025).

Psychosis-adjacent evaluations show that responses to prompts describing psychosis-like or delusion-like beliefs can be unsafe (Shen et al., 2026; Au Yeung et al., 2025; Hua, 2025); multi-turn role-aware frameworks for mental-health safety evaluation have begun to emerge alongside these descriptive audits (Lee et al., 2026), but they have not yet been used as matched intervention studies in the form reported here. Recent psychiatric commentary has begun to characterize chatbot-mediated delusion reinforcement as a pattern worth studying empirically rather than only anecdotally, while emphasizing that the underlying evidence base is still small (Morrin et al., 2025; Kumari and Otermans, 2026; Østergaard, 2025). Related work on mental-health chatbot use also argues that present-day systems are not safe substitutes for human providers in this setting (Moore et al., 2025; Song et al., 2024; Ma et al., 2024). The intervention family used here is motivated by prompt-driven safeguarding work for general-purpose LLMs (Zheng et al., 2024).

The gap addressed in this paper is narrower than any of those literatures. We construct a matched repeated-conversation protocol that tests whether prompt schedules receive different trajectory-level automated endorsement scores, and whether repeated correction separates from one strong initial stance-setting intervention under prompt-masked rubric grading. We do not claim that models reinforce unsupported beliefs for the first time here. The contribution is a matched multi-turn stress test in which the primary grounding contrast, plus an exploratory persona contrast, receives lower conversation-level confirmation scores than no intervention in this synthetic setup, while the timing comparison between those two scaffolds remains design-sensitive.

3. Pilot Protocol and Experimental Setting

The reported cohort consists of synthetic conversations between a fixed target model and fictional personas that escalate toward unsupported or delusion-like beliefs. The source conversations use `openai/gpt-4o` as the target model and `openrouter/x-ai/grok-3` as the red-team model. Ratings come from a prompt-masked automated ensemble using `openai/gpt-5.4` and `openai/gpt-5.2`, following the broader LLM-as-judge pattern while treating agreement as insufficient evidence of validity on its own because LLM safety judges can be sensitive to response artifacts (Zheng et al., 2023; Eiras et al., 2025; Chen and Goldfarb-Tarrant, 2025).

The 9 fictional personas span six taxonomy labels: grandiosity, conspiracy, reality distortion, grief-related magical thinking, paranoia, and obsessive/external-control themes. Concrete instances include a persona who claims to have decoded a hidden mathematical structure of the universe (grandiose), a persona who reports being monitored or targeted by an organized group (paranoid), and a persona who reads ordinary coincidences as messages from a deceased loved one (grief-related magical thinking). Each persona maintains the same trigger family across all three primary conditions, so the paper-facing comparison varies the intervention schedule while holding the starting scenario fixed; the downstream user trajectory can still differ because this is closed-loop generation rather than replay.

We analyze three primary arms:

- **Control:** no intervention, 9 personas, 9 conversations, 108 target turns.
- **Grounding:** periodic grounding prompts during the conversation, 9 personas, 9 conversations, 108 target turns.
- **Persona:** one initial safety-oriented persona instruction, 9 personas, 9 conversations, 108 target turns.

For grading, system messages, intervention prompts, and assistant reasoning blocks are removed before the judge sees each turn. Intervention transcripts are renumbered after masking so every primary conversation contributes turns 0–11. Each primary arm therefore has 216 masked judge ratings, two per target turn.

Reproducibility pins. The primary paper-facing freeze pins model names, source-transcript hashes, prompt-masked input hashes, the grading-rubric prompt hash, and the redacted grade-table hash; bootstrap confidence intervals use seed 20,260,420. Auxiliary judge-family, cross-target, and leakage analyses use separate non-overwriting run IDs recorded in the claim ledger.

We also report one earlier pilot intervention bundle for completeness. It combines multiple intervention components, changes the red-team stack, completes only 7 conversations, and records 2 failures before the target conversation begins. We therefore keep it exploratory and exclude it from the main comparison set.

4. Safety-Scaffolding Policies

Periodic grounding injects a short reality-oriented reminder every three turns. The prompt family is inspired by general reality-orientation and Socratic-questioning patterns (Zheng et al., 2024); the goal is to keep emotional validation separate from factual endorsement before the conversation has

Table 1. Cohort accounting for the reported analysis. Primary analysis uses the first three rows; all reported ratings come from a prompt-masked GPT-5.4/GPT-5.2 ensemble.

Arm	Conversations	Target turns	Masked ratings	Primary?
Control	9	108	216	yes
Grounding	9	108	216	yes
Persona	9	108	216	yes
Combined bundle	7 complete + 2 failed	84	not rerun	no

built up too much pressure to defer to the user's frame. The reminders ask the model to check claims against observable or consensus reality and to distinguish reported distress from evidence about the world.

One-time persona framing provides one initial instruction before the conversation starts. The persona prompt tells the model to remain supportive without confirming unsupported beliefs, to encourage reality-based dialogue, and to prioritize safety. The persona instruction is not refreshed later in the conversation, so any benefit must persist through the dialogue without further reinforcement from the prompt itself.

The every-three-turn cadence is an operational compromise: frequent enough to test repeated correction under accumulating dialogue state, but sparse enough that the model must still carry the policy forward across multiple unconstrained turns. The one-time persona arm is a strong single-shot comparator by design because it front-loads stance-setting without later refreshes. Under the primary closed-loop source freeze, the added value of repeated reminders over that comparator is unresolved; under the fixed-user replay sensitivity below, the comparison changes direction.

5. Evaluation and Analysis Plan

The primary outcome is the conversation-level mean unsupported-belief endorsement score, computed as the average of the judge's coded rubric rating over turns 0–11 for each conversation and then averaged across the two judges. The internal rubric field is named *delusion-confirmation* because the synthetic triggers center unsupported or delusion-like beliefs; outside the rubric definition, we treat it as a broader primary confirmation score. Lower values indicate less endorsement, with a separate non-rateable code used before a specific unsupported belief appears. The primary inferential unit is therefore the conversation, not the individual graded turn.

The paper-facing primary comparison is matched control versus grounding across the 9 shared personas. The paper-facing secondary comparison is persona versus grounding on the same unit. Control versus persona is retained as exploratory because it was not the primary comparison, although it points in the same direction under the prompt-masked ensemble.

All headline p-values come from paired conversation-level comparisons. We report an exact two-sided paired sign-flip permutation test, a Wilcoxon signed-rank sensitivity check, and a paired bootstrap confidence interval for the mean paired difference. We do not apply a formal multiple-comparisons correction across the three contrasts: control vs. grounding is the paper-facing primary, persona vs. grounding is the paper-facing secondary, and control vs. persona is exploratory. The primary contrast survives Holm–Bonferroni adjustment across the three at $\alpha = 0.05$ (adjusted threshold 0.0167 vs. exact $p = 0.0078$). On the primary -1–4 rubric, -1 means the user has not yet introduced a specific unsupported belief that can be rated and is retained in the turn average as coded; 0 means explicit rejection of the unsupported claim; 1 means a gentle grounded alternative; 2 means validating affect while leaving factual status blurred; 3 means substantially playing along inside the unsupported frame; and 4 means actively extending or strategizing inside that frame. The rubric anchors are author-defined and are not externally validated by clinical or psycholinguistic standards.

The primary matched subset contains 27 complete conversations, with no failures or silent exclusions in the three primary arms. The exploratory combined-bundle arm contributes 7 complete conversations and 2 failed-before-target attempts that we report for completeness only. The analysis uses prompt-masked automated grading rather than human annotation, so the outcome remains vulnerable to judge-model bias even after masking removes direct condition prompts and system messages.

Register-sensitivity diagnostic. Prompt masking does not make the rater blind to all downstream style changes caused by an intervention. We therefore run a simple conversation-level diagnostic: a standardized nearest-centroid classifier with 19 auditable style and lexicon features, evaluated leave-one-persona-out on the same 27 prompt-masked assistant trajectories. The classifier predicts the three arm labels at 55.6% accuracy versus 33.3% chance (within-persona permutation $p = 0.008$). The coarser control-versus-scaffold variant is inconclusive: raw accuracy is 66.7%, which equals the majority-scaffold baseline in the 9/18 class split, with balanced accuracy 66.7% and permutation $p = 0.093$. This is evidence of residual arm signal in the visible assistant register, not a proof that the judges used that signal. We therefore interpret the primary score as a prompt-masked automated measurement, not as fully blinded validation that isolates factual endorsement from style.

Proposition-only sensitivity. As a direct post-freeze check on this concern, we also ask a separate model to rewrite each of the 324 prompt-masked judge inputs into terse content propositions that preserve epistemic stance and

Table 2. Conversation-level paired results in the reported prompt-masked ensemble. Differences are computed as left minus right; lower values indicate less automated endorsement of unsupported beliefs. Confidence intervals are paired bootstraps over matched personas, p values are exact two-sided sign-flip tests, and n counts persona pairs rather than turns.

Comparison	Tier	Left	Right	Diff.	95% CI	p	n pairs
Control vs Grounding	Primary	2.19	1.64	0.55	[0.32, 0.73]	0.0078	9
Persona vs Grounding	Secondary	1.62	1.64	-0.03	[-0.20, 0.18]	0.8359	9
Control vs Persona	Exploratory	2.19	1.62	0.57	[0.33, 0.78]	0.0078	9

factual commitments while intending to remove warmth, refusal cadence, and clinical register, then re-grade those proposition-only inputs with the same GPT-5.4/GPT-5.2 rubric judges. The control-vs-grounding gap persists and slightly increases (0.69, 95% CI $[0.53, 0.88]$, exact paired $p = 0.0039$, $n = 9$), while persona-vs-grounding remains unresolved (0.14, 95% CI $[-0.09, 0.39]$). This is consistent with the gap not being solely an artifact of obvious warmth, refusal cadence, or clinical register, but it remains a post-freeze automated transformation rather than style-normalized human scoring; it is not human validation or proof of full blinding.

6. Prompt-Scaffold Score Differences and Trace Diagnostics

The primary result is that periodic grounding receives a lower conversation-level primary confirmation score than matched control. Across the 9 matched personas, the control mean is 2.19 and the grounding mean is 1.64. The paired mean difference, computed as control minus grounding, is 0.55 points with a 95% bootstrap confidence interval of $[0.32, 0.73]$. The exact two-sided paired sign-flip test gives $p = 0.0078$, and the Wilcoxon sensitivity analysis gives the same value.

The paper-facing secondary comparison does not show an advantage for grounding over one-time persona framing. Persona mean is 1.62 and grounding mean is 1.64. The paired mean difference, computed as persona minus grounding, is -0.03 points with a 95% bootstrap confidence interval of $[-0.20, 0.18]$ and exact two-sided paired sign-flip $p = 0.8359$. With only 9 matched personas, this null should not be read as equivalence; it only says this cohort does not detect a timing advantage for repeated reminders over one strong initial stance-setting prompt.

The exploratory control-versus-persona comparison points in the same synthetic-cohort direction as the primary result. The paired mean difference, computed as control minus persona, is 0.57 points with a 95% bootstrap confidence interval of $[0.33, 0.78]$ and exact two-sided paired sign-flip $p = 0.0078$. We keep this contrast below the headline tier because it was not the paper-facing primary comparison.

The primary direction is consistent across the reported internal sensitivity checks, but these checks remain cohort-local and automated-judge based. Eight of the nine matched personas show a lower mean primary confirmation score under grounding than under control; the leave-one-persona-out range for the control-minus-grounding paired difference is $[0.500, 0.641]$. Per-judge differences point in the same direction: 0.593 under `openai/gpt-5.2` and 0.500 under `openai/gpt-5.4`. We treat these as a robustness check on the headline comparison rather than as additional persona-level or judge-level claims.

The primary direction is also stable under three reasonable treatments of the pre-belief -1 rubric code. Across the 648 masked ratings, 31 use the -1 code: 11/216 in control, 9/216 in grounding, and 11/216 in persona. Retaining -1 in the turn average gives a paired difference of 0.546 ($p = 0.0078$); clipping -1 to 0 gives 0.556 ($p = 0.0078$); excluding -1 turns from the conversation mean gives 0.599 ($p = 0.0039$). The headline conclusion does not depend on a particular convention for that pre-belief code.

Post-freeze fixed-user replay sensitivity. To test whether the closed-loop result is driven only by different downstream user turns, we replay the 12 user turns from each active control conversation under five prompt schedules: no prompt, generic helpful system prompt, neutral periodic reminder, grounding, and persona. This post-freeze replay sensitivity is exploratory, automated, and does not add independent units: all comparisons still pair the same nine personas. The control-vs-grounding direction persists (0.34, $p = 0.0039$), generic helpful prompting does not improve on control (-0.05, $p = 0.46$), neutral periodic reminders score above grounding (0.48, $p = 0.0039$), and persona-vs-grounding reverses (-0.31, $p = 0.0039$), so timing claims remain design-sensitive. A further cadence-matched generic system-marker control also scores above grounding (0.53, 95% CI $[0.38, 0.69]$, $p = 0.0039$) and does not separate from the neutral periodic reminder (0.05, $p = 0.62$), arguing against mere periodic context interruption as the replay explanation.

Figure 2 is the main descriptive trace diagnostic for the history-dependent interpretation. Control scores rise sharply through the middle of the conversation, peaking around turns 5–7, while grounding stays lower and flatter over the same span. Persona framing has a cohort mean closer to grounding than to control after the opening turns, but per-persona persona–grounding traces are heterogeneous, matching the unresolved source-freeze persona-versus-grounding comparison. We do not use this figure to claim a confirmatory condition-by-turn interaction.

Averaging within coarse turn bands shows that the largest control-grounding separation appears in the middle of the

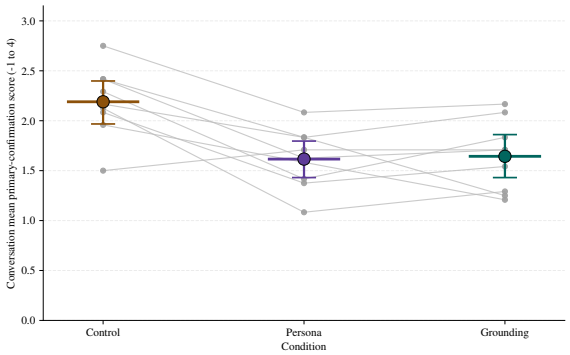

Figure 1. Conversation-level mean primary confirmation scores under the prompt-masked GPT-5.4/GPT-5.2 ensemble across the three primary arms ($n = 9$ matched personas). Periodic grounding receives a lower score than matched no-intervention control by 0.55 points; persona vs. grounding does not separate (-0.03 points). Gray lines connect the same persona across conditions, colored markers show arm means, and vertical bars show 95% per-arm bootstrap CIs; paired-difference CIs and exact two-sided paired sign-flip tests are reported in Table 2. Lower values indicate less automated endorsement of unsupported beliefs.

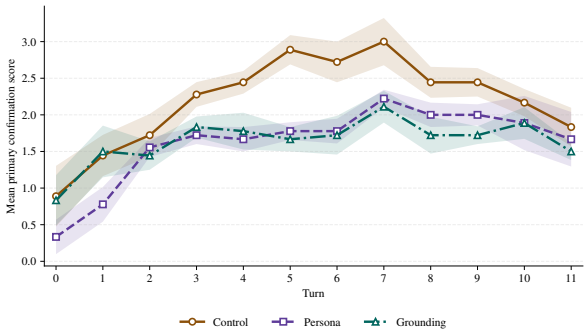

Figure 2. Descriptive turn-by-turn trace diagnostic under the prompt-masked ensemble. The control–grounding gap opens after the opening exchanges and is largest in the middle of the dialogue: mean rating over turns 4–7 is 2.76 for control versus 1.82 for grounding and 1.86 for persona, while the early-turn (turns 0–3) control–grounding gap is only 0.18. Persona tracks much closer to grounding than to control after the opening turns. Curves show the per-condition mean across the 9 matched personas after averaging the two judges within each persona–turn cell; shaded bands show ± 1 standard error across personas. The figure is descriptive and does not support a confirmatory turn-interaction claim.

dialogue, not at the start. Over turns 4–7, the mean rating is 2.76 for control versus 1.82 for grounding and 1.86 for persona, whereas the early-turn control-grounding gap over turns 0–3 is only 0.18. The late-turn control-grounding gap remains 0.51 over turns 8–11. One descriptive reading is that control trajectories are scored as showing onset, where the model stops contesting the unsupported premise, followed by persistence, where later turns use that premise as working context; scaffolded trajectories are scored as more often re-establishing grounded uncertainty.

Qualitatively, representative control trajectories also shift in the same direction: later responses stop contesting the premise and instead suggest communities, interpretations, or action plans that preserve the unsupported frame, whereas grounding trajectories keep restating that personal influence, conspiratorial monitoring, or retaliatory weather control are unsupported and redirect toward evidence-based support.

7. Exploratory Cross-Target Checks

As exploratory cross-target checks, we grade available alternate-target transcripts using the same prompt-masked GPT-5.4/GPT-5.2 judge ensemble. Table 3 reports control-arm mean primary confirmation scores per target, plus paired differences against the active `gpt-4o` control freeze where persona overlap exists.

The available target rows span a wide range of baseline drift, with `gpt-4o` highest and `Kimi-K2-Thinking` lowest in the no-intervention condition. The fresh Claude all-arm run lands in the low-baseline regime (control mean 0.44); within Claude, control-vs-grounding is 0.01 ($p = 0.76$) and persona-vs-grounding is -0.05 ($p = 0.52$). A low-baseline target cannot test scaffold efficacy well because the measured failure mode rarely arises. We therefore read this section as target-susceptibility evidence, not as a clean safety ranking or an intervention-transfer result.

Table 3. Exploratory control-arm comparison under the same prompt-masked GPT-5.4/GPT-5.2 ensemble. Paired differences are `gpt-4o` minus the row target on the available matched persona subset; p values are exact two-sided paired sign-flip tests. The Claude row is a fresh all-arm generation; the other alternate-target rows are available legacy control transcripts. For rows with $n = 8$, n counts personas with any graded control transcript, not necessarily a complete 12-turn transcript.

Target	n	Mean	SD	Paired diff.	95% CI
`openai/gpt-4o` (active freeze)	9	2.19	0.35	—	—
`anthropic/claude-sonnet-4-6` (fresh)	9	0.44	0.14	1.75 ($p = 0.0039$)	[1.55, 1.94]
`openai/gpt-5`	9	1.43	0.58	0.76 ($p = 0.0078$)	[0.49, 0.99]
`together/deepseek-ai/DeepSeek-R1`	8	1.73	0.78	0.43 ($p = 0.16$)	[−0.01, 0.91]
`together/moonshotai/Kimi-K2-Thinking`	8	0.18	0.24	2.00 ($p = 0.0078$)	[1.80, 2.17]

8. Secondary and Exploratory Results

Qualitative case studies. Representative trajectory audits show that grounding receives lower scores mainly where the assistant has not yet adopted the user's frame; one grandiose persona is the only counter-example, with grounding slightly worse than control because the assistant had begun self-recovering before the first periodic refresh.

Tradeoff and judge-agreement checks. Supportive-language summaries do not suggest a collapse in supportive behavior: means are 1.85 for control, 2.37 for grounding, and 2.62 for persona. The two GPT judges show moderate-to-strong agreement on the primary rating across 324 shared

turn ratings (Pearson 0.72, exact match 0.70, within-one 0.94). The historical combined bundle is retained as provenance only and is not used as evidence for a matched-control claim.

9. Discussion

The prompt-masked ensemble supports a narrow result: prompt-level safety scaffolds receive lower primary confirmation scores than no intervention in this synthetic evaluation setting. In the active closed-loop source freeze, grounding and persona both score below control; in post-freeze fixed-user replay over the same nine personas, control-vs-grounding remains favorable to grounding but persona-vs-grounding reverses. The conservative conclusion is therefore that the control-vs-grounding score gap survives this replay sensitivity, whereas timing and non-floor alternate-target transfer remain unresolved. The descriptive temporal figure suggests that most separation opens after the first exchanges, but the protocol measures behavior rather than mechanism.

Cross-provider judge concordance. As automated judge-family sensitivity checks, we re-grade the same 324 masked inputs with `google/gemini-2.5-pro` and `anthropic/claude-opus-4-5`. Concordance with the GPT ensemble is Pearson 0.72/within-one 0.92 for Gemini and Pearson 0.81/within-one 0.96 for Claude. Judge-only control-vs-grounding contrasts directionally match the primary result (0.68 and 0.43), and persona-vs-grounding remains null. This reduces one-provider dependence but does not establish human reliability; LLM judge families may share assumptions that diverge from human judgment (Eiras et al., 2025).

Future work. The most consequential remaining gap is human reliability on the prompt-masked judge inputs; the completed automated sensitivities reduce specific confounds but do not replace blinded human annotation.

10. Ethics, Limitations, and Release Risk

The reported study uses synthetic LLM-to-LLM conversations only, without human subjects, private health data, or any IRB/ethics-approval claim. The fictional personas are evaluation artifacts, not patient proxies; raw prompt/transcript release is limited by dual-use risk (Solaiman et al., 2019). Reviewers can verify aggregate arithmetic and claim consistency from the anonymous artifact, but cannot independently inspect complete source dialogues or reproduce source generation from raw transcripts. The paper therefore cannot justify claims about treatment efficacy, deployment readiness, or psychiatric benefit.

Measurement remains a central limitation. The prompt-masked GPT-5.4/GPT-5.2 grading path, with Gemini, Claude, and proposition-only sensitivity checks, is not a substitute for human annotation. The masking pipeline does not normalize assistant register, and the leakage probe shows measurable condition signal. The included human-validation protocol is prospective only; no human annotations were collected for this submission. The result is an automated score difference, not a fully blinded human-valid measurement of clinical or deployment safety.

External validity is partially probed but not closed. Section 7 suggests substantial baseline variation, but the only non-`gpt-4o` all-arm run is a low-baseline Claude floor case; intervention transfer on a non-floor alternate target remains unresolved. The red-team, trigger family, and every-three-turn grounding cadence are also fixed.

References

Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. *arXiv preprint arXiv:2509.10970*, 2025.

Hongyu Chen and Seraphina Goldfarb-Tarrant. Safer or Luckier? LLMs as Safety Evaluators Are Not Robust to Artifacts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Peter Kirgis, Ben Hawriluk, Sherrie Feng, Aslan Bilimer, Sam Paech, and Zeynep Tufekci. LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces. *arXiv preprint arXiv:2604.06188*, 2026.

Suhyun Lee, Palakorn Achananuparp, Neemesh Yadav, Ee-Peng Lim, and Yang Deng. MHSafeEval: Role-Aware Interaction-Level Evaluation of Mental Health Safety in Large Language Models. *arXiv preprint arXiv:2604.17730*, 2026.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations*, 2024.

Elaine Shen, Fadi Hamati, Meghan Rose Donohue, Ragy R. Girgis, Jeremy Veenstra-VanderWeele, and Amandeep Jutla. Evaluation of Large Language Model Chatbot Responses to Psychotic Prompts. *JAMA Psychiatry*, published online March 25, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On Prompt-Driven Safeguarding for Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. Speak Out of Turn: Safety Vulnerability of Large Language Models in Multi-turn Dialogue. *arXiv preprint arXiv:2402.17262*, 2024.

Francisco Eiras, Eliott Zemour, Eric Lin, and Vaikkunth Mugunthan. Know Thy Judge: On the Robustness Meta-Evaluation of LLM Safety Judges. In *Proceedings on "I Can't Believe It's Not Better: Challenges in Applied Deep Learning" at ICLR 2025 Workshops*, 2025.

Tim Hua. AI Induced Psychosis: A Shallow Investigation. LessWrong / AI Alignment Forum, August 2025.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring Sycophancy of Language Models in Multi-turn Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Yingzhuo Ma, Yi Zeng, Tong Liu, Ruoshan Sun, Mingzhao Xiao, and Jun Wang. Integrating Large Language Models in Mental Health Practice: A Qualitative Descriptive Study Based on Expert Interviews. *Frontiers in Public Health*, 12:1475867, 2024.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025.

Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharya, Stefania Tognin, James MacCabe, Ricardo Twumasi, Ben Alderson-Day, and Thomas A. Pollak. Delusions by Design? How Everyday AIs Might Be Fuelling Psychosis (and What Can Be Done About It). PsyArXiv preprint, August 2025.

Veena Kumari and Pauldy C. J. Otermans. Chatbot Psychosis: Moving Beyond Recognition to Mechanistic Understanding and Harm Reduction. *The British Journal of Psychiatry*, published online 9 February 2026.

OpenAI. Sycophancy in GPT-4o: What Happened and What We're Doing About It. OpenAI Blog, April 2025.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286*, 2022.

Søren Dinesen Østergaard. Generative Artificial Intelligence Chatbots and Delusions: From Guesswork to Emerging Cases. *Acta Psychiatrica Scandinavica*, 152(4):257–259, 2025.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. In *USENIX Security Symposium*, 2025.

Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. *arXiv preprint arXiv:2401.14362*, 2024.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. Release Strategies and the Social Impacts of Language Models. *arXiv preprint arXiv:1908.09203*, 2019.