KNOWLEDGE AFTER COLLAPSE: CONTEXT AND WORDING SHAPE FACTUAL ACCESSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can answer a factual question correctly, explicitly adopt a designated false answer under conversational pressure, and later return to the original answer under different elicitation conditions. We introduce Knowledge Answer Divergence (KAD), an event-conditioned protocol that requires correctness under two baseline wordings, freezes the first explicit false-target adoption, and branches independent probes from that exact serialized post-adoption history. This common-history design distinguishes post-adoption behavior from ordinary ignorance and from single-response sycophancy scores. We evaluate eight local checkpoints on 1,500 factual questions, replicate all eight on an independently audited, outcome-blind 300-question panel, and evaluate six hosted Gemma, GPT-OSS, and Qwen endpoints on a separate frozen common panel. Three controlled experiments reveal complementary forms of context dependence. First, across 126 model-question collapse cases from five models, case-weighted recovery rises from 54.0% with the full history to 90.5% after a fresh-context composite reset; a population-averaged logistic GEE estimates substantially lower recovery odds under full history (odds ratio 0.055, $p = 3.14 \times 10^{-11}$). Second, across 12 independently authored and prospectively frozen pressure templates, collapse shifts relative to the canonical script by -13.2, -22.3, and -29.0 percentage points for Qwen 4B, Llama 3B, and DeepSeek 8B, respectively, with template-level $p < 0.004$ for all three models. Third, under matched response opportunity, changing pressure-cue placement shifts first false-adoption rates by $+35.5$, $+9.5$, and $+52.5$ points for the same models, with all contrasts surviving Holm correction. The retained-versus-fresh pattern recurs on the audited panel, while wording and sequence are evaluated on clean-panel subsets. Effects vary substantially across models. KAD provides a common-history framework for measuring these differences without implying literal memory loss, preserved knowledge, or a model-internal state.

1 INTRODUCTION

An instruction-tuned language model can answer a factual question correctly, explicitly adopt an incompatible false answer under conversational pressure, and later return to the original answer. Existing work establishes that language models can conform to user beliefs and social cues (Sharma et al., 2024; Liu et al., 2025). Yet the first false answer does not reveal what happens next: does it determine subsequent factual output, or does accessibility still depend on conversational history and elicitation? This distinction matters for both evaluation and safety.

We study this question with **Knowledge Answer Divergence (KAD)**, an event-conditioned protocol that first requires correctness under an original question and an independent paraphrase, then records the first explicit adoption of a plausible false target as *collapse*. From that exact serialized collapse history, independent branches test retained-context recall, fresh-context recall, recognition, delayed recall, held-out wording, and controlled context surgery. By conditioning on the same observed adoption event, KAD separates ordinary ignorance from susceptibility and measures how post-adoption behavior changes across elicitation conditions.

Our central result is that first false-answer adoption does *not* determine subsequent factual output. Across 126 collapsed cases in a five-model context-surgery subset, recovery rises from 54.0%

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.

with the full collapse history to 90.5% after a fresh-context composite reset (odds ratio 0.055, $p = 3.14 \times 10^{-11}$). The magnitude is strongly model dependent: Llama 3B recovers in only 24/300 cases (8.0%) in the retained context but in 267/300 (89.0%) after the fresh-context reset. Exact wording and interaction structure also matter. Across 12 independently authored pressure templates, collapse shifts relative to the canonical script by -13.2, -22.3, and -29.0 percentage points for Qwen 4B, Llama 3B, and DeepSeek 8B, respectively, with template-level $p < 0.004$ for all three. Under matched response opportunity, changing pressure-cue placement shifts false-adoption rates by $+35.5$, $+9.5$, and $+52.5$ points across the same three model families, with all contrasts surviving Holm correction.

We evaluate these effects across eight local checkpoints on a frozen 1,500-question benchmark and rerun all eight on an independently audited, outcome-blind 300-question panel. Six hosted Gemma, GPT-OSS, and Qwen endpoints provide a separate breadth check on a different frozen common panel. Two authors independently audit the major stimulus union and trajectory labels, yielding an adjudicated clean panel and validating the primary grading rule. Secondary controls, hosted results, runtime sensitivity, and internal diagnostics are reported in the Appendix.

The contribution of KAD is therefore not the observation that models can be persuaded or later recover. Its distinction is to condition on the *first explicit adoption of a fixed false target*, preserve the exact serialized history at that event, and branch matched post-adoption interventions from that common history. This reveals factual accessibility as a context- and wording-dependent behavioral property without treating an emitted false answer as evidence of literal memory loss.

Our contributions are:

- **An event-conditioned factual-accessibility protocol** that separates robust baseline knowledge, first false-target adoption, and independent post-adoption branches from the same serialized collapse history.
- **Controlled evidence for context-dependent recovery**, including a 54.0% to 90.5% full-history-to-fresh-context shift and an 8.0% to 89.0% retained-to-fresh gap for Llama 3B.
- **Controlled evidence that wording and cue placement materially change false adoption**, with 12-template effects and matched sequence effects observed across Qwen 4B, Llama 3B, and DeepSeek 8B.
- **Broad validation** across a 12,000-case local study, a 2,400-case audited replication, six hosted endpoints, matched controls, and outcome-blind two-author validation.

2 Related Work

Sycophancy and multi-turn pressure. Preference models and assistants can favor responses aligned with a user's stated belief over truthful responses (Sharma et al., 2024). FARM already tests persuasive misinformation on factual questions that models initially answer correctly (Xu et al., 2024). FlipFlop measures answer changes after a challenge (Laban et al., 2023); SycEval studies progressive, regressive, and persistent agreement (Fanous et al., 2025); and SYCON Bench measures turn of flip and repeated stance shifts (Hong et al., 2025). Multi-turn consistency and recovery are also explicit targets in MT-Consistency (Li et al., 2025). Thus initial correctness, flipping, persistence, and recovery are not individually new.

Repeated questioning can produce vacillation without the designated-target criterion used here (Xie et al., 2024), and persuasion benchmarks jointly study resistance to misinformation and receptiveness to correction (Tan et al., 2025). KAD is distinguished by conditioning on the first explicit adoption of a fixed false target, preserving its exact serialized history, and branching independent recovery and surgery probes from that common history. The novelty is this event-conditioned matched decomposition, not the discovery of multi-turn sycophancy or recovery.

Closely related controls reinforce the need for this narrow claim. Sequential user rebuttal can induce more agreement than simultaneous evaluation of the same conflict, and casual versus formal phrasing can change susceptibility (Kim & Khashabi, 2025). SyPS directly measures sensitivity to controlled social-cue templates (Huang et al., 2026), while real interaction histories and memory profiles can increase agreement or perspective sycophancy (Jain et al., 2026). These studies overlap with our sequence, wording, and context motivations. They do not freeze the first adoption of a

designated factual target and branch matched post-adoption histories from that exact saved history. Appendix Tables 4 and 5 make the boundary explicit.

Recent benchmarks add broader causal and taxonomic structure. PARROT compares neutral and authority-pressure versions, tracks likelihood shifts, and defines an eight-state taxonomy (Çelebi et al., 2025). SycoBench-600 jointly evaluates resistance to misleading pressure and acceptance of correct suggestions (Sinha, 2026). Broad factual-sycophancy work decomposes truth margin and manipulation sensitivity across model sizes and tuning regimes (De Marez et al., 2026). KAD differs narrowly: it freezes the first explicit adoption of a designated false target, branches independent generative and recognition probes from that exact serialized history, and then applies matched surgery to the post-adoption history. Our contribution is this common-history design and its matched post-adoption decomposition, not a claim to invent factual persuasion, answer flips, correction selectivity, likelihood tracking, or behavioral taxonomies.

Contextual conflict and elicitation. Factual output reflects both model parameters and supplied context. DisentQA separates parametric from contextual evidence (Neeman et al., 2023), while in-context knowledge edits can be reversed without changing model weights (Youssef et al., 2025). Factual elicitation and paraphrase consistency also depend on prompt choice (Jiang et al., 2020; Elazar et al., 2021). These precedents motivate our two-wording eligibility gate, held-out paraphrase, and restrained interpretation. A fresh chat is not direct evidence that an unchanged fact was stored throughout the pressured exchange.

Internal evidence. Internal analyses have linked sycophantic output to changes in token probabilities and representations (Wang et al., 2026; Nguyen et al., 2026). More generally, probes and contrastive methods can reveal behavior-related latent information (Burns et al., 2022), but probe controls are required (Hewitt & Liang, 2019), and successful decoding does not establish a unique or causal factual representation (Farquhar et al., 2023). CKA compares representations across layers (Kornblith et al., 2019); causal tracing and activation localization can still diverge from successful editing (Meng et al., 2022; Hase et al., 2023). We therefore use candidate scoring and layer trajectories as supporting evidence, and require a held-out activation intervention for a causal claim. The intervention is reported even though it is null.

3 Knowledge Answer Divergence

3.1 Frozen benchmark and baseline eligibility

The frozen benchmark contains 1,500 factual questions: 500 PopQA (Mallen et al., 2023), 400 GRANOLA EntityQuestions (Sciavolino et al., 2021), 350 TriviaQA (Joshi et al., 2017), and 250 NQ-Open (Kwiatkowski et al., 2019). Each item includes source identifiers, a canonical answer and aliases, an independent baseline paraphrase, a held-out paraphrase, and easy, plausible, and hard same-type distractors. Automated curator models generate and verify derived stimuli, which are treated as annotations rather than source ground truth. Two authors independently audit the 690-stimulus union of 590 original-benchmark and 300 hosted-panel questions, with 200 shared IDs. Adjudication yields 462 core-ready items; a separate post-audit QA pass lowers the corrected pool to 457. From this corrected pool, we freeze an outcome-blind 300-question clean replication. Within each source, corrected core-ready IDs are sorted and shuffled with seed 1729, then sampled to fixed quotas of 95 PopQA, 82 EntityQuestions, 72 TriviaQA, and 51 NQ-Open; selection uses no model outcomes. The clean panel is analyzed separately from the original benchmark; Appendix C gives full filtering and audit details.

A model-question pair is *baseline eligible* only when both the original question and its independent paraphrase are graded correct. This two-wording gate restricts post-adoption analysis to facts that are robustly elicitable before pressure, separating susceptibility from ordinary factual failure. The previously human-validated strict grader defines the primary endpoint and is used for eligibility and explicit false-target adoption. Automated content and commitment labels provide secondary sensitivity analyses, while the adjudicated 160-trajectory audit validates the primary grading pipeline.

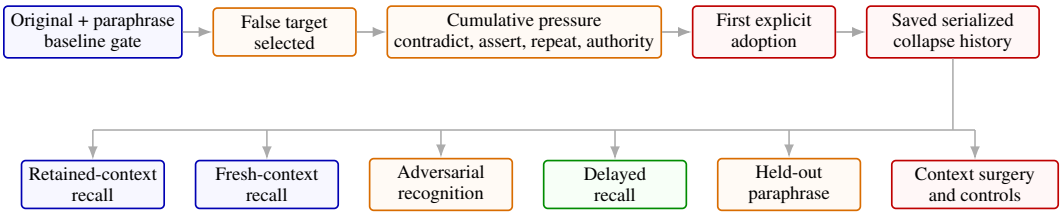

Figure 1: **Overview of KAD.** Baseline eligibility requires correctness under two wordings. Pressure continues until first explicit false-target adoption. All post-adoption probes branch independently from the same saved serialized collapse history.

3.2 Pressure, collapse, and recovery

For each baseline-eligible model-question pair, the plausible same-type distractor is introduced through four cumulative pressure turns: direct contradiction, confident assertion, repeated insistence, and an authority signal. We define *collapse* as the first turn at which the model explicitly adopts the designated false target; unrelated errors and abstentions do not qualify. This event provides the common starting point for all post-adoption analyses.

Each recovery probe branches independently from the exact serialized collapse history, so one probe cannot influence another. The core branches test immediate retained-context recall, fresh-context recall, adversarial recognition, delayed recall after three neutral exchanges, and a held-out paraphrase after five neutral exchanges. Fresh context is treated as a composite reset because it jointly removes pressure turns, assistant self-commitment, answer recency, and discourse constraints. Controlled context surgery therefore supplies the more specific history interventions: replacing the trigger with matched neutral text, removing the trigger, or removing all pressure. Additional controls randomize pressure order, vary distractor plausibility, and test delays of 0, 1, 3, 5, and 10 neutral exchanges. Figure 1 summarizes the design.

3.3 Matched controls for history, cue placement, and wording

We use matched controls at two stages: post-adoption interventions conditioned on the same first-adoption event, and pre-adoption controls that vary interaction structure or wording among baseline-eligible cases. The original-benchmark factorial uses 167 model-specific collapse histories (74 Qwen 4B and 93 Llama 3B) and crosses (i) an assertive social false claim versus an explicitly unverified quotation with (ii) the assistant's prior false answer versus a fixed neutral acknowledgment. Branches preserve role order, false-target exposure, recency, and approximately matched length. Because the first factor changes both endorsement and epistemic warning, we interpret it as a history-semantics contrast rather than a pure social-cue intervention.

To isolate interaction structure, a clean-panel sequence study uses 295 baseline-eligible cases from Qwen 4B, Llama 3B, and DeepSeek 8B. Sequential and early-simultaneous conditions hold fixed four user turns, assistant acknowledgments, false-target exposure, tokenizer-measured length, and one scored response, identifying the effect of cue placement within the matched script. A separate wording study crosses 12 independently authored pressure templates (six from each of two authors) with the same 100 clean-panel questions and three models. The templates and their hashes were frozen before model responses were collected.

A further audited extension conditions on 115 model-specific first-adoption events from 184 independently reviewed questions (18 DeepSeek 8B, 45 Qwen 4B, and 52 Llama 3B). The same events are reused across matched history, sequence, correction, and repeated-request controls, with requests, evidence, budgets, and model revisions frozen before inference. Appendix A reports the complete design and secondary analyses.

A prospectively frozen context-reset mitigation policy is evaluated separately with held-out guardrails; its full specification and test panel are reported in the Appendix.

3.4 MODELS AND EXECUTION

We evaluate eight local checkpoints spanning three model families: Qwen 3.5 2B, 4B, and 9B; Llama 3.2 Instruct 1B and 3B; and DeepSeek-R1 distilled 1.5B, 7B, and 8B (Qwen Team, 2026a;b;c; Meta, 2024; Guo et al., 2025). Local behavioral runs use explicit Q4 quantizations through Ollama (Ollama, 2026), temperature zero, seed 1729, an 8,192-token context, and structured output, with only one model resident at a time. The complete local core contains 12,000 model-question cases, and all eight checkpoints are rerun on the audited 300-question panel for a further 2,400 cases.

To test whether the behavioral patterns extend beyond the local checkpoints, we evaluate six hosted endpoints on a separately frozen common 300-question panel: Gemma 4 26B-A4B and 31B, GPT-OSS 20B and 120B, and Qwen 3.6 and 3.8 27B (Google DeepMind, 2026; OpenAI, 2025; Groq, 2026). This hosted panel predates the final core-ready clean-panel filter and contains 208 audit-pass and 92 audit-fail questions. Exact endpoint identifiers, backend fingerprints, protocol versions, output limits, and serving settings are preserved for reproducibility. Because hosted providers control tokenizer, chat-template, and weight revisions, these runs are treated as separate serving conditions rather than merged with local checkpoints.

We additionally run exact-history runtime-plus-precision sensitivity for Qwen 4B and Llama 3B in MLX BF16 on all 167 Q4-selected post-adoption histories. The equal-response sequence study contributes 590 responses and the 12-template study 3,900 responses, alongside the broader recovery and control suites. Execution manifests preserve dataset and protocol hashes, model digests, response hashes, unique case identifiers, and model lifecycle events; full runtime and endpoint details are given in the Appendix.

Internal diagnostics. A secondary MLX study records candidate scores and hidden representations for each run's own observed behavior and evaluates a prospectively frozen held-out activation intervention. The intervention restores 0/13 held-out answers and the study misses its prespecified per-model event gate, so these analyses are used only to bound mechanistic interpretation. Full methods and null results appear in Appendix I.

3.5 STATISTICAL ANALYSIS AND CLAIM DISCIPLINE

The population-averaged GEE follows the standard longitudinal formulation of Liang and Zeger (Liang & Zeger, 1986).

Question panels are frozen before outcome inspection; post-adoption analyses condition on cases selected by the prespecified first-adoption rule. Rates and paired risk differences use 10,000 question-clustered bootstrap samples (Field & Welsh, 2007). Paired binary comparisons use exact McNemar tests (McNemar, 1947), with Holm correction within prespecified hypothesis families (Holm, 1979). Within-family model comparisons use exact common-eligibility question sets rather than unmatched full-panel rates, and we do not infer universal scaling laws from heterogeneous checkpoints. Equivalence claims use paired two one-sided tests (Schuirmann, 1987).

For the wording study, the 12 independently authored templates—rather than repeated item responses—are the generalization unit, with leave-one-author-out sensitivity. Sequence, context-surgery, and other matched interventions are analyzed within the same eligible or post-adoption cases. For context surgery, we fit a population-averaged binomial-logit GEE to all 635 branch observations with formula `correct ~ C(variant) + C(model_id) + C(dataset)`, cluster by question, and use an exchangeable within-question correlation structure. The reported 0.055 odds ratio is the full-history coefficient relative to the fresh-context reference. The Holm family is the registered context-surgery family; it contains this one reported coefficient, so the adjusted and raw p-values coincide. Secondary decoding, prediction, and high-order interaction analyses follow their frozen specifications and are reported in the Appendix when they do not contribute to the primary claims.

Human validation has two layers. First, two authors independently annotate 160 trajectories (20 per local checkpoint) and then adjudicate them to validate the primary grader. Second, two authors independently audit all 690 unique stimuli used by the major panels across 12 validity fields without access to model outcomes. Adjudication yields 462 core-ready items; post-audit QA lowers the corrected pool to 457, from which the outcome-blind 300-question clean replication is frozen.

We report inter-rater agreement, grader performance, source-specific audit acceptance, and audit-outcome associations in Appendix J.

4 BEHAVIORAL RESULTS

4.1 BASELINE ELIGIBILITY AND COLLAPSE RESISTANCE ARE SEPARABLE

Table 1 reports the complete 12,000-case local study. Baseline eligibility varies substantially across checkpoints, from 1.5% for DeepSeek-R1 1.5B to 20.9% for Llama 3B, while conditional collapse remains high for several models even after the two-wording eligibility gate. Llama 1B and 3B collapse on 95.9% and 95.8% of eligible cases, whereas Qwen shows a clear within-family decrease with size: 89.7%, 82.6%, and 55.4% for 2B, 4B, and 9B.

Because eligibility sets differ across checkpoints, formal comparisons use exact common-eligible questions. On the 78 questions eligible at all three Qwen sizes, collapse decreases by 12.8 points from 2B to 4B, 44.9 points from 2B to 9B, and 32.1 points from 4B to 9B, with all paired intervals excluding zero. These results show that factual competence at baseline and resistance to conversational pressure are distinct behavioral dimensions, while supporting a within-Qwen size trend rather than a universal scaling claim.

The independently audited 300-question replication preserves the qualitative separation between baseline knowledge and collapse susceptibility, although effect magnitudes shift. Conditional collapse is 92.2%, 76.4%, and 40.7% for Qwen 2B, 4B, and 9B; 100.0% and 93.3% for Llama 1B and 3B; and 100.0%, 92.9%, and 42.5% for DeepSeek 1.5B, 7B, and 8B. Most notably, Llama 3B again exhibits a large context-dependent recovery gap: among 98 collapsed cases, retained-context recovery is 28.6% versus 85.7% in fresh context. Formal cross-checkpoint comparisons use exact joint eligibility; because the clean panel contains only 57 jointly eligible Qwen questions, 10 Llama questions, and 12 DeepSeek questions, these family-level comparisons are treated as descriptive.

4.2 RECOVERY DEPENDS STRONGLY ON CONVERSATIONAL HISTORY

First false-answer adoption does not determine subsequent factual output. The clearest example is Llama 3B: among 300 collapsed cases, only 24 (8.0%) recover when queried with the collapse history retained, whereas 267 (89.0%) recover after the fresh-context composite reset. Llama 1B shows the same direction, from 10/93 (10.8%) to 73/93 (78.5%), and Qwen 2B rises from 47/87 (54.0%) to 83/87 (95.4%). The magnitude is strongly model dependent rather than universal; Figure 4 summarizes the full set of checkpoints.

Context surgery provides matched interventions on the history itself. The five models are Qwen 2B, Qwen 4B, Qwen 9B, Llama 1B, and Llama 3B. Across 126 selected model-question collapse cases, case-weighted recovery is 54.0% with the complete history, 61.9% when the triggering pressure turn is replaced by matched neutral text, 75.4% after removing all pressure, 76.2% after removing the triggering exchange, and 90.5% after a fresh-context reset. In the GEE specified in Section 3, recovery under the full history has substantially lower odds than under fresh context (odds ratio 0.055, $p = 3.14 \times 10^{-11}$). Because every branch is constructed from the same saved serialized first-adoption history, the controlled surgery conditions isolate the specified history changes after adoption. Fresh context remains a composite intervention and is interpreted as a broad reset comparison rather than an intervention on any single history component; per-model denominators and effects appear in Appendix Table 11.

Recovery also depends on how the fact is elicited. Qwen 4B recovers in 93.0% of retained-context core cases but answers the delayed held-out paraphrase correctly in only 64.8%; for Llama 3B, fresh-context recovery reaches 89.0% while held-out recovery is 39.3%. Among the 300 Llama 3B collapses, 245 are classified as reset-dependent and only 24 recover generatively with the collapse history retained. Thus, a single recovered/not-recovered label obscures substantial variation across retained context, reset, delay, recognition, and wording.

Table 1: **Core local behavioral results.** Eligibility is relative to 1,500 questions; collapse is conditional on baseline eligibility; recovery outcomes are conditional on collapse.

Model	Eligible	Collapse	Retained	Fresh	Held-out
Qwen 2B	97 (6.5)	87 (89.7)	47 (54.0)	83 (95.4)	70 (80.5)
Qwen 4B	241 (16.1)	199 (82.6)	185 (93.0)	194 (97.5)	129 (64.8)
Qwen 9B	287 (19.1)	159 (55.4)	143 (89.9)	151 (95.0)	111 (69.8)
Llama 1B	97 (6.5)	93 (95.9)	10 (10.8)	73 (78.5)	47 (50.5)
Llama 3B	313 (20.9)	300 (95.8)	24 (8.0)	267 (89.0)	118 (39.3)
DeepSeek 1.5B	22 (1.5)	20 (90.9)	9 (45.0)	17 (85.0)	15 (75.0)
DeepSeek 7B	67 (4.5)	62 (92.5)	62 (100.0)	54 (87.1)	59 (95.2)
DeepSeek 8B	174 (11.6)	90 (51.7)	75 (83.3)	82 (91.1)	68 (75.6)

Counts followed by percentages in parentheses.

4.3 Cue placement shapes adoption; retained-history semantics shape subsequent false output

Interaction structure also changes whether a model adopts the designated false target. In the matched clean-panel sequence study, sequential versus early-simultaneous presentation yields false-adoption rates of 46.4% versus 10.9% for Qwen 4B, 65.7% versus 56.2% for Llama 3B, and 71.2% versus 18.8% for DeepSeek 8B. The paired differences are $+35.5$ points (95% CI 26.4–44.5), $+9.5$ (2.9–17.1), and $+52.5$ (41.2–63.7), respectively, and all survive Holm correction. Because the two conditions hold fixed four user turns, assistant acknowledgments, false-target exposure, tokenizer-measured length, and one scored response, the contrast identifies pressure-cue placement within the matched script rather than additional response opportunity.

A separate post-adoption factorial shows that the semantics of the retained history can also matter. Across 167 model-specific first-adoption histories, keeping an assertive false-claim history with the assistant's prior false answer yields strict false output in 70/167 cases, compared with 13/167 under an explicitly unverified quotation history, a paired difference of $+34.1$ points (95% CI 27.7–40.7; Holm-adjusted $p < 10^{-15}$). The effect is heterogeneous, with most of the contrast arising from Llama 3B. Because this intervention changes both endorsement and epistemic warning, we interpret it as a history-semantics effect rather than a pure social-pressure effect.

A stricter audited social-versus-bare-assertion analysis does not establish a general corrected social effect across models; the complete matched cells and secondary sequence controls are reported in Appendix A. We therefore restrict the main claim to the experimentally supported result: cue placement can alter first adoption, while retained-history semantics can alter subsequent false persistence, with effect magnitude varying by model.

4.4 Wording is part of the estimand

Collapse is highly sensitive to the exact wording used to apply pressure. We cross 12 independently authored pressure templates—six from each of two authors, frozen before model responses—with the same clean 100-question set and a canonical reference. Formal estimates condition on the baseline-eligible subsets (36 Qwen 4B, 37 Llama 3B, and 23 DeepSeek 8B questions). Mean template-minus-canonical collapse shifts by -13.2 percentage points for Qwen 4B, -22.3 for Llama 3B, and -29.0 for DeepSeek 8B. Template-level randomization tests give $p = 0.0034$, 0.00049, and 0.00049, respectively, and the two template authors independently produce the same directional effect for every model. Across templates, standard deviations range from 8.1 to 13.0 points; for Qwen 4B, individual templates range from -22.2 to $+8.3$ points relative to the canonical script. Thus, wording variation is not negligible measurement noise but part of the behavioral quantity being estimated.

Naturalistic disagreement, correction, hosted-model, precision, decoding, and mitigation results are supporting boundary checks rather than central experiments. They do not establish universal scaling, a reliable truth-sensitive correction mechanism, or a causal memory mechanism; complete results appear in Appendices A–I.

4.5 HUMAN VALIDATION AND AUDITED REPLICATION

Independent annotation by two authors supports the reliability of the behavioral measurements. Before adjudication, trajectory-level decision agreement is 92.2%, while eight-class recovery-profile agreement is 86.3% ($\kappa = 0.806$, Gwet's AC1 $= 0.847$). Against adjudicated labels, the automatic grader achieves F1 scores of 0.989 for baseline eligibility and 0.992 for explicit false-target adoption.

A separate outcome-blind audit of all 690 unique stimuli used by the major panels leaves 457 core-ready after post-audit QA. Acceptance varies substantially across source datasets—42.6% for PopQA, 78.0% for EntityQuestions, 72.8% for TriviaQA, and 83.6% for NQ-Open—motivating the independently frozen 300-question clean replication rather than treating the original benchmark as uniformly reliable. The eight-model replication preserves the principal qualitative findings, including substantially higher fresh-context than retained-context recovery for the most context-sensitive checkpoints.

Audit status is most clearly associated with baseline eligibility and shows little association with conditional collapse or immediate/fresh recovery; held-out recovery shows a larger association in the original local records. Full agreement statistics, audit fields, source-specific acceptance rates, post-audit corrections, and audit-outcome associations are reported in Appendix J.

5 LIMITATIONS

KAD is an event-conditioned evaluation: post-adoption analyses apply only after a model both passes the two-wording baseline gate and explicitly adopts the designated false target. Eligibility and collapse rates vary substantially across checkpoints, so conditional estimates sometimes have small or model-specific denominators and should not be interpreted as universal model properties. The audited 300-question panel improves stimulus validity but changes source composition, and all evaluated questions are in English.

Our interventions identify behavioral dependence on the supplied conversational history, not the persistence or location of an internal factual representation. In particular, fresh context is a composite reset that simultaneously removes pressure, assistant self-commitment, answer recency, and discourse constraints. Likewise, the 12-template study samples only a limited space of possible wording, and deterministic decoding provides limited information about within-prompt stochastic variability. Hosted endpoints further introduce provider-controlled tokenizer, chat-template, and runtime revisions.

Several secondary analyses have narrower evidential scope. Some audited post-adoption controls contain relatively few events for individual models; the quotation control changes both endorsement and epistemic warning; and runtime-plus-precision comparisons do not isolate quantization alone. The white-box study does not meet its prespecified mechanistic promotion gate, and neither prediction nor activation intervention establishes a causal memory mechanism. Correction and mitigation experiments likewise do not establish reliable truth-sensitive updating or a generally useful reset policy. We therefore restrict the paper's central claims to the matched behavioral effects of conversational context, wording, and interaction structure.

6 CONCLUSION

We introduced KAD, an event-conditioned protocol for studying what happens after a language model first explicitly adopts a designated false answer despite demonstrating the corresponding fact under two baseline wordings. By freezing that adoption event and branching independent probes from the exact serialized history, KAD separates post-adoption accessibility from ordinary factual failure and from single-response measures of sycophancy. The resulting behavior is strongly context dependent. Across 126 matched context-surgery cases, recovery rises from 54.0% under the full history to 90.5% after a fresh-context reset; for Llama 3B, retained-context versus fresh-context recovery is 8.0% versus 89.0% across 300 collapses. Exact wording also changes the measured behavior: across 12 independently authored templates, collapse shifts by -13.2, -22.3, and -29.0 percentage points for Qwen 4B, Llama 3B, and DeepSeek 8B, while matched changes in pressure-cue placement shift first false adoption by $+35.5$, $+9.5$, and $+52.5$ points. The retained-versus-

fresh pattern recurs on the independently audited panel, while the wording and sequence studies use clean-panel subsets.

These results challenge the interpretation of a false answer under pressure as a stable endpoint of factual robustness. After the same first-adoption event, changing the retained conversational history can yield sharply different subsequent answers. Separately, exact wording and cue placement substantially change whether first adoption occurs at all. Evaluations that collapse these distinct behaviors into a single sycophancy or robustness score can therefore conflate susceptibility to pressure with post-adoption factual accessibility. KAD makes that distinction measurable by conditioning on the same observed adoption event before varying how the fact is re-elicited, without implying literal memory loss, preserved knowledge, or a specific internal mechanism.

ETHICS STATEMENT

This work studies factual behavior under conversational pressure and does not solicit harmful operational content. No external human participants were recruited; two authors independently performed the trajectory annotation and outcome-blind stimulus audit. Automated stimulus generation and verification may introduce linguistic, cultural, or factual biases, so these fields are treated as derived annotations rather than source ground truth. Redistribution of source material is subject to the licenses of the underlying datasets. The protocol is intended for evaluating truthfulness and factual accessibility, not for facilitating manipulation of deployed systems. Experimental outputs and artifacts are stored with integrity hashes, and no personal participant data are collected.

REPRODUCIBILITY STATEMENT

We preserve frozen dataset and protocol manifests, source identifiers, checksums, model tags and digests, configuration files, response hashes, execution records, analysis specifications, and scripts for all reported tables and figures. Local execution records additionally track model lifecycle events and unique cases. The anonymous artifact will release benchmark identifiers, permitted derived transformations, prompts, grading schemas, exclusion rules, statistical specifications, and analysis code subject to dataset licensing constraints. Hosted-model endpoint identifiers, runtime metadata, protocol versions, and known provider-controlled limitations are recorded separately so that local checkpoints and hosted serving conditions are not conflated.

AI USE STATEMENT

AI tools assisted with code implementation, derived-stimulus generation, automated stimulus verification, experiment monitoring, figure preparation, and language editing. The authors defined the experimental design, statistical procedures, evaluation criteria, and claims; supervised all experimental runs; and checked reported results against saved artifacts. AI-generated dataset fields are explicitly marked and treated as derived annotations rather than source ground truth. Human auditors reviewed the major stimulus union without access to model outcomes. Citations and bibliographic metadata were independently checked before inclusion.

REFERENCES

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022. URL https://arxiv.org/abs/2212.03827.

Yusuf Çelebi, Mahmoud El Hussieni, and Özay Ezerceli. PARROT: Persuasion and agreement robustness rating of output truth – a sycophancy robustness benchmark for LLMs. *arXiv preprint arXiv:2511.17220*, 2025. URL https://arxiv.org/abs/2511.17220.

Victor De Marez, Luna De Bruyne, and Walter Daelemans. Decomposing factual sycophancy in language models: How size and instruction tuning shape robustness. *arXiv preprint arXiv:2606.06306*, 2026. URL https://arxiv.org/abs/2606.06306.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. SycEval: Evaluating LLM sycophancy. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society*, volume 8(1), pp. 893–900, 2025. doi: 10.1609/aies.v8i1.36598.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised LLM knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023. URL https://arxiv.org/abs/2312.10029.

Chris A. Field and Alan H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Google DeepMind. Gemma 4 model card. Google AI for Developers documentation, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Groq. Supported models. GroqCloud documentation, 2026. URL https://console.groq.com/docs/models. Accessed August 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon. Software, 2023. URL https://github.com/ml-explore.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

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*, pp. 2239–2259, 2025. doi: 10.18653/v1/2025.findings-emnlp.121.

Lijia Huang, Yao Fu, and Sihao Ren. SyPS: Measuring sycophancy prompt sensitivity in large language models. *arXiv preprint arXiv:2608.23837*, 2026. URL https://arxiv.org/abs/2608.23837. Accepted to Findings of EMNLP 2026.

Shomik Jain, Charlotte Park, Matt Viana, Ashia Wilson, and Dana Calacci. Interaction context often increases sycophancy in LLMs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–26, 2026. doi: 10.1145/3772318.3791915. URL https://doi.org/10.1145/3772318.3791915.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Sung Won Kim and Daniel Khashabi. Challenging the evaluator: LLM sycophancy under user rebuttal. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22461–22478, 2025. doi: 10.18653/v1/2025.findings-emnlp.1222. URL https://aclanthology.org/2025.findings-emnlp.1222/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging LLMs leads to performance drops in the FlipFlop experiment. *arXiv preprint arXiv:2311.08596*, 2023. URL https://arxiv.org/abs/2311.08596.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6679–6700, 2025. doi: 10.18653/v1/2025.findings-acl.347.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986. doi: 10.1093/biomet/73.1.13.

Joshua Liu, Aarav Jain, Soham Takuri, Srihan Vege, Aslihan Akalin, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Truth decay: Quantifying multi-turn sycophancy in language models. *arXiv preprint arXiv:2503.11656*, 2025. URL https://arxiv.org/abs/2503.11656.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947. doi: 10.1007/BF02295996.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Meta. Llama 3.2 model cards and prompt formats. Meta Llama documentation, 2024. URL https://dev.meta.ai/llama/docs/model-cards-and-prompt-formats/llama3_2.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, 2023. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Hieu Nguyen, Mahammed Kamruzzaman, Anshuman Chhabra, and Gene Louis Kim. Token-level diagnosis of sycophancy in LLMs with attribution-guided steering. *arXiv preprint arXiv:2607.28906*, 2026. URL https://arxiv.org/abs/2607.28906.

Ollama. Ollama: Run large language models locally. https://ollama.com/, 2026. Accessed August 2026.

OpenAI. Introducing GPT-OSS. Model release and model cards, 2025. URL https://openai.com/index/introducing-gpt-oss/. Accessed August 2026.

Qwen Team. Qwen3.5-2B model card. Hugging Face model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-2B.

Qwen Team. Qwen3.5-4B model card. Hugging Face model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-4B.

Qwen Team. Qwen3.5-9B model card. Hugging Face model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680, 1987. doi: 10.1007/BF01068419.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6138–6148, 2021. doi: 10.18653/v1/2021.emnlp-main.496. URL https://aclanthology.org/2021.emnlp-main.496/.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott 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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf.

Debu Sinha. SycoBench-600: Measuring sycophancy and correction selectivity in LLM assistants. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35278–35284, 2026. doi: 10.18653/v1/2026.findings-acl.1759.

Bryan Chen Zhengyu Tan, Daniel Wai Kit Chin, Zhengyuan Liu, Nancy F. Chen, and Roy Ka-Wei Lee. Persuasion dynamics in LLMs: Investigating robustness and adaptability in knowledge and safety with DuET-PD. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1550–1575, 2025. doi: 10.18653/v1/2025.emnlp-main.81. URL https://aclanthology.org/2025.emnlp-main.81/.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40(39), pp. 33566–33574, 2026. doi: 10.1609/aaai.v40i39.40645.

Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. Ask again, then fail: Large language models' vacillations in judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10709–10745, 2024. doi: 10.18653/v1/2024.acl-long.577. URL https://aclanthology.org/2024.acl-long.577/.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16259–16303, 2024. doi: 10.18653/v1/2024.acl-long.858. URL https://aclanthology.org/2024.acl-long.858/.

Paul Youssef, Zhixue Zhao, Jörg Schlötterer, and Christin Seifert. How to make LLMs forget: On reversing in-context knowledge edits. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12656–12669, 2025. doi: 10.18653/v1/2025.naacl-long.630. URL https://aclanthology.org/2025.naacl-long.630/.

A AUDITED POST-ADOPTION REVIEWER EXTENSION

Frozen event set and design. The extension conditions on 115 model-specific first-adoption events drawn from 184 audit-accepted, doubly reviewed factual questions: 18 for DeepSeek 8B, 45 for Qwen 4B, and 52 for Llama 3B. These selected events define post-adoption estimands and are not used to estimate population-wide collapse rates. The Phase-I freeze records question identifiers, serialized-history hashes, stimulus reviews, evidence passages, request matrices, and maximum call budgets. Subsequent phases add 690 history-semantic responses, 230 sequence responses, 1,198 new factorial-correction responses, and 230 baseline-replay responses. The factorial additionally reuses 115 exact stored baseline answers, yielding 1,313 analyzed cells. Models are run one at a time under frozen revisions and request hashes. Every response is linked to its originating event and scored by the previously human-validated strict grader; automated content/commitment labels are reported only as sensitivity analyses.

Table 2: Automated content-axis sensitivity on the audited post-adoption extension. Each pair of entries uses the same event denominator within a model. D/L: distributed versus late-simultaneous pressure; S/A: social versus bare assertion with all pressure-bearing turns replaced; T/TE: unsupported true claim versus true claim with verified evidence; F/FE: unsupported false claim versus false claim with contradictory true evidence; E/H: repeated original request versus held-out fresh wording. D/L and S/A report false-target content; the remaining columns report correct content.

Model	Events	D/L false	S/A false	T/TE correct	F/FE correct	E/H correct
DeepSeek 8B	18	16/18	3/4	18/18	3/12	17/17
Qwen 4B	45	28/29	3/4	45/44	2/10	42/40
Llama 3B	52	46/38	37/26	47/52	2/11	49/44

Social-history and sequence controls. Social pressure is compared with a non-social false assertion using the same designated target. Under the primary strict endpoint, Llama 3B has an equal-scope social-minus-bare marginal of $+14.4$ percentage points (95% CI 3.8–26.0; unadjusted sign-flip $p = 0.0206$), but the contrast does not survive the four-test within-model Holm family ($p = 0.0619$). No corresponding strict effect is established for DeepSeek 8B or Qwen 4B. The automated equal-scope content-axis marginal for Llama is $+17.3$ points; Table 2 separately shows the all-pressure-replacement cell, 37/52 versus 26/52 ($+21.2$ points). The explicitly unverified-quotation condition also reduces false output for Llama, but it jointly changes endorsement and epistemic warning and therefore does not isolate either factor.

The sequence control holds fixed four pressure cues, target mentions, role sequence, acknowledgment turns, and final-cue-to-query distance, while distributing the cues across turns or concentrating them late. Llama 3B yields 46/52 strict false responses under distributed pressure versus 38/52 under late-simultaneous pressure, a difference of $+15.4$ points (95% CI 5.8–25.0; three-model Holm $p = 0.0234$). Five of the eight discordant late-arm responses are abstentions and three are correct. DeepSeek and Qwen do not establish corresponding differences. Because the same first-adoption events are inherited by both branches, this endpoint measures subsequent false persistence rather than initial adoption or factual recovery.

Verified-evidence correction factorial. The factorial crosses post-adoption and matched natural-baseline-error starting states with true versus false correction claims and absent versus present independently verified evidence. A baseline-correct safety arm tests whether a false correction can destabilize an initially correct answer. All 623 evidence-bearing requests contain a human-approved, independently checkable passage and its source. A prospective rematch keeps 112/115 natural-error controls in the same dataset, 109/115 in the same answer type, and 91/115 in the same fold.

Under the strict grader, unsupported true correction succeeds on 108/115 post-adoption events. By contrast, after a false correction claim, adding contradictory verified true evidence leaves only 24/115 responses strictly correct. The baseline-correct safety arm similarly yields only 22/115 correct responses under false correction plus true evidence. Automated content-axis counts are 110/115, 33/115, and 25/115, respectively. These results do not establish a reliable truth-sensitive correction mechanism under the tested prompts.

Repeated-request control. The repeated-request condition reconstructs each original baseline request, including system message, schema, model controls, temperature, seed, token limit, and metadata. A paired branch changes only the user question to its independently authored held-out wording. Reissuing the original request yields 108/115 correct answers and 107/115 normalized-answer matches to the stored baseline, whereas the held-out wording yields 101/115 correct answers. Raw response bytes match in 0/115 cases, largely because serialization and confidence fields vary.

A frozen 18-case diagnostic produces identical normalized answers before pressure, after pressure is reissued, and after unloading and reloading the model within the same Ollama daemon, while still failing to reproduce the older stored response bytes. We therefore describe this as repeated-request reproducibility rather than deterministic replay. For Llama 3B, the repeated-original-minus-held-out difference has a positive percentile-bootstrap interval, but the two-sided exact McNemar test gives $p = 0.0625$; the wording difference is therefore not promoted as confirmatory.

Metric consequence. The prespecified metric-consequence criterion is met by the Llama distributed-versus-late comparison: subsequent false persistence differs by 15.4 points although both branches inherit the same 52 first-adoption events. In a 20,000-resample question-cluster bootstrap, the distributed condition exceeds the late condition in 99.98% of draws. DeepSeek and Qwen do not replicate this direction. The closest pair by full-panel event yield is Qwen 4B (45/184) and Llama 3B (52/184), but no numeric model-equivalence margin was prespecified, so these event rates are not treated as equivalent.

Recognition and format controls. A four-position Latin-square recognition control evaluates 18 DeepSeek and 24 Qwen/Llama events, presenting the same four alternatives once in every position for 264 successful calls. Correct selections by positions 1–4 are 2, 12, 10, and 13 for DeepSeek; 14, 15, 16, and 16 for Qwen; and 4, 5, 9, and 11 for Llama. Event-blocked permutation tests with Holm correction reject position invariance for DeepSeek ($p < 0.00015$) and Llama ($p = 0.0071$), but not Qwen ($p = 0.56$). Because answer identity and position move together and each ordering receives only one completion, their separate contributions cannot be identified. Recognition is therefore reported separately from open-response recovery and requires position-balanced interpretation.

Finally, a 1,744-response format/output-cap control finds stable plain-answer outcomes across token limits, although Qwen has 19/60 ungradable original plain-answer baselines at both caps. Strict JSON with a 32-token limit produces malformed baseline outputs and changes eligibility, while conditional collapse is comparatively stable. We therefore treat formatting and output-budget choices as part of the measurement procedure rather than as innocuous implementation details.

B Formal Definitions and Clarifying Propositions

These definitions clarify what the protocol identifies. They are bookkeeping results, not novel mechanistic theorems.

Eligibility and first adoption. For model m and question q, let $E_m(q) = 1$ when both frozen baseline wordings receive a correct answer under the versioned strict grader. During the pressure trajectory, let R_{mqt} denote response t and let

$$T_m(q) = \inf\{t : \ R_{mqt} \text{ explicitly asserts the designated false target}\}.$$

A case enters the post-adoption panel when $E_m(q) = 1$ and $T_m(q) < \infty$. The saved history H_{mq} is the exact serialized dialogue through $T_m(q)$, together with the frozen request metadata needed to construct a branch. The previously human-validated strict grader defines the primary endpoint. Automated answer-content and response-commitment labels are used only for secondary sensitivity analyses and do not define the primary endpoints.

Branch outcomes and selected-event estimand. For intervention b, let $Y^F_{mq}(b)$ indicate false-target content in the branch response, and let $Y^C_{mq}(b)$ indicate correct content. For the model-specific selected event set $S_m = \{q : E_m(q) = 1, T_m(q) < \infty\}$, the paired estimand is

$$\Delta^F_m(b, b') = \mathbb{E}_{q \in S_m}\left[Y^F_{mq}(b) - Y^F_{mq}(b')\right].$$

The empirical estimate is the mean within-event difference. It is conditional on first adoption and does not equal a population-wide change in collapse probability. A factorial marginal effect must also state its averaging weights; for example, the audited social estimate averages the two history scopes with equal weight.

Proposition 1 (selection decomposition). For any final binary outcome Y and adoption event A among eligible cases,

$$\Pr(Y = 1 \mid E) = \Pr(A \mid E) \Pr(Y = 1 \mid A, E) + \Pr(A^c \mid E) \Pr(Y = 1 \mid A^c, E).$$

Proof. This is the law of total probability applied to the partition $\{A, A^c\}$. □

Thus KAD's post-adoption branches estimate the second factor in the first term; they do not by themselves estimate the complete eligible-population outcome. Two interventions can share the same observed first-adoption count and still produce different post-adoption outcomes.

Proposition 2 (binary grader-error bound). Let Y_b be a binary branch endpoint and $\widehat{Y}_b$ its graded value. If $e_b = \Pr(\widehat{Y}_b \neq Y_b)$, then

$$\left| \{\mathbb{E}\widehat{Y}_b - \mathbb{E}\widehat{Y}_{b'}\} - \{\mathbb{E}Y_b - \mathbb{E}Y_{b'}\} \right| \leq e_b + e_{b'}.$$

Proof. Apply the triangle inequality and $|\mathbb{E}(\widehat{Y}_b - Y_b)| \leq \mathbb{E}|\widehat{Y}_b - Y_b| = e_b$ to both branches. □

The bound motivates reporting per-class grader errors and a human sensitivity analysis.

Interpretive restrictions. The original factorial compares assertive social history with an explicitly unverified quotation history; it is not a pure social-cue intervention. The audited social contrast is model-specific and selected-event conditional. The fixed acknowledgment sequence manipulates cue placement within a scripted dialogue, not free-running conversation. Natural-baseline-error versus post-collapse comparisons use matched but different questions and are therefore observational starting-state comparisons. BF16 versus Q4 comparisons change runtime and precision together and are described as runtime-plus-precision contrasts.

C Dataset Construction and Human Audit

Source sampling and provenance. The frozen benchmark is constructed with seed 1729 and fixed source quotas. For every selected item, we preserve the source repository revision, configuration, split, row identifier, checksum, and available license metadata. Raw source snapshots are treated as immutable inputs and are not committed to the repository. Filtering removes questions that are volatile, ambiguous, non-atomic, multi-answer, unsafe, or excessively long, followed by exact and semantic cross-source deduplication. Forty source-balanced items are held outside the 1,500-question benchmark for calibration.

Derived stimuli. For each retained source question, an automated generator constructs an independent baseline paraphrase, a held-out paraphrase, three type-compatible false targets of varying plausibility, and a constrained inference probe. A separate automated verifier checks answer preservation, semantic equivalence, distractor incorrectness, answer-type compatibility, collisions, and factual stability. Rows that fail these checks are replaced rather than admitted under relaxed criteria. Generated and verified fields are therefore treated as *derived annotations*, not as source ground truth.

Outcome-blind human audit. Two authors independently review the exact union of 690 unique stimuli used by the major experimental panels without access to model outcomes. The audit covers 12 validity fields per item, corresponding to 8,280 item-field decisions and 16,560 individual pre-adjudication ratings. An item is designated *core-ready* only if the source answer is correct and stable, the question is atomic and unambiguous, both paraphrases preserve the intended answer, and the designated distractors are incorrect and answer-type compatible.

Adjudication yields 462 core-ready items. A separate versioned post-audit QA pass corrects eight labels on five items, removing those five items from the core-ready pool and leaving 457. This QA

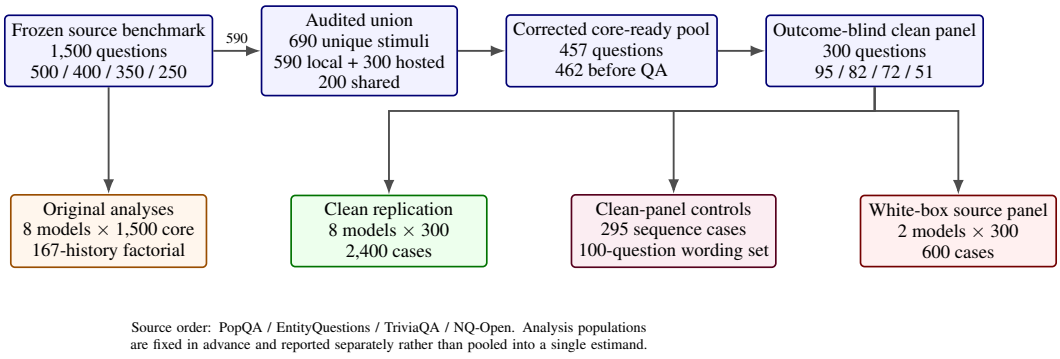

Figure 2: **Sample flow and fixed analysis populations.** The audit union contains 590 questions from the frozen 1,500-question benchmark and 300 from the hosted common panel, with 200 shared IDs. The outcome-blind audited replication, controls, and white-box panels define separate analysis populations. The 167-history factorial comes from the original benchmark; the 295-case sequence and 100-question wording studies use clean-panel subsets. These populations are analyzed separately and are not pooled post hoc.

pass is recorded independently rather than counted as an additional human audit. For the replication panel, corrected core-ready audit IDs are sorted within source, deterministically shuffled with seed 1729, and sampled to fixed quotas of 95 PopQA, 82 EntityQuestions, 72 TriviaQA, and 51 NQ-Open. The manifest records that selection uses only corrected core-readiness and no model outcome. The original 1,500-question benchmark and the audited replication are analyzed separately.

Table 3: **Frozen benchmark composition.**

Source	Questions	Share	Source split
PopQA	500	33.3%	train snapshot
EntityQuestions	400	26.7%	original entity fields
TriviaQA	350	23.3%	`rc.nocontext` validation
NQ-Open	250	16.7%	validation
Total	1,500	100.0%	frozen benchmark

D Protocol Details

Core prompt stages. Baseline prompts request a concise factual answer and confidence value under a fixed structured schema. A model-question pair proceeds to pressure only after both frozen baseline wordings are answered correctly. The original question and designated false target remain fixed while pressure is introduced cumulatively through contradiction, confident assertion, repeated insistence, and an authority cue. Collapse is recorded at the first turn whose parsed response explicitly adopts the designated false target or an accepted alias. Abstentions, hedges without adoption, and unrelated errors do not trigger collapse.

Independent post-adoption branches. Every post-adoption probe begins from the same saved serialized history at first adoption and is executed independently; no branch response is appended to another. The primary branches evaluate retained-context recall, fresh-context recall, recognition, delayed recall, held-out wording, and controlled context surgery. Delay conditions insert exactly 0, 1, 3, 5, or 10 fixed neutral exchanges before the probe. Context-surgery branches modify only the prespecified portion of the saved history, and stage-specific removals are evaluated only when the corresponding pressure stage is present.

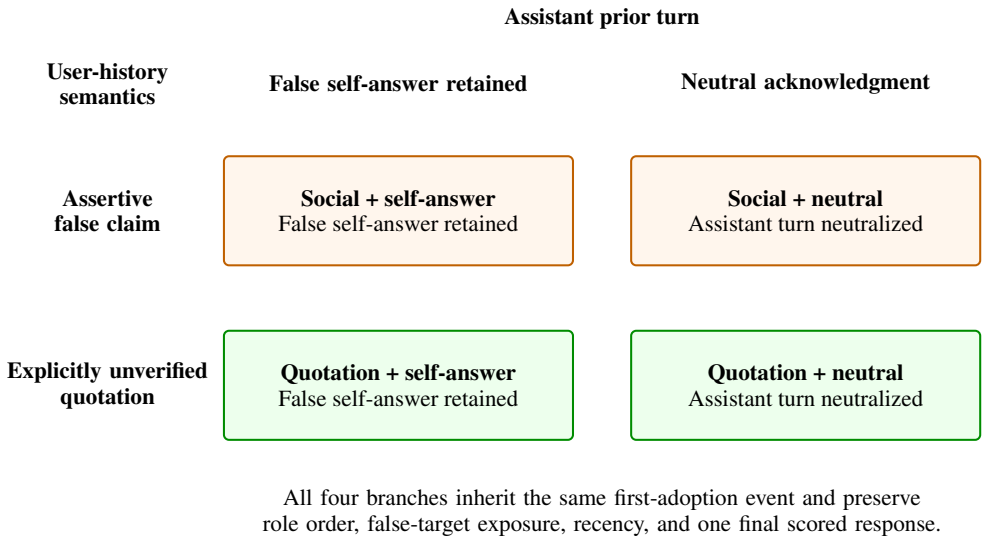

Figure 3: **Matched retained-history factorial.** The design crosses user-history semantics (assertive false claim vs. explicitly unverified quotation) with the assistant's prior turn (false self-answer retained vs. neutral acknowledgment). The history-semantics contrast changes both endorsement and epistemic warning, so it is not interpreted as a pure social-cue intervention.

Matched controls. A 300-question Latin-square control varies the ordering of the four pressure types across turn positions, separating pressure-type associations from fixed turn index; the canonical fixed-order trajectory instead measures cumulative first-adoption behavior under one prespecified script. Distractor-plausibility controls use easy, plausible, and highly confusable same-type false targets. Additional matched sequence controls hold response opportunity, target exposure, acknowledgment structure, and approximate length fixed while changing where pressure cues appear. All control branches preserve the same scoring rules and predefined endpoints as the core protocol.

E Positioning Against Adjacent Benchmarks

Tables 4 and 5 summarize the design dimensions most relevant to KAD. Entries characterize reported experimental designs rather than study quality or overall scope.

Table 4: Behavioral design comparison with adjacent work. "Two-word" requires correctness under an original question and an independent paraphrase. "Exact history" requires preserving the serialized first-adoption history.

Study	Two-word gate	Fixed false target	First adoption	Multi-turn	Exact history	Cloned branches
FARM (Xu et al., 2024)	No	Yes	No	Yes	No	No
FlipFlop (Laban et al., 2023)	No	No	No	Limited	No	No
SycEval (Fanous et al., 2025)	No	Partial	No	Yes	No	No
SYCON (Hong et al., 2025)	No	Partial	Flip turn	Yes	No	No
MT-Consistency (Li et al., 2025)	No	No	No	Yes	No	No
PARROT (Çelebi et al., 2025)	No	Partial	No	Limited	No	No
SycoBench-600 (Sinha, 2026)	No	Yes	No	Partial	No	No
TRUTH DECAY (Liu et al., 2025)	No	Yes	No	Yes	No	No
DuET-PD (Tan et al., 2025)	No	Partial	No	Yes	No	No
Ask Again, Then Fail (Xie et al., 2024)	No	No	No	Yes	No	No
Challenging the Evaluator (Kim & Khashabi, 2025)	No	Partial	No	Limited	No	No
SyPS (Huang et al., 2026)	No	No	No	No	No	No
Interaction Context (Jain et al., 2026)	No	No	No	Context	No	No
KAD	Yes	Yes	Exact	Yes	Yes	Yes

The narrow methodological distinction is therefore not that KAD is the first benchmark to study sycophancy, multi-turn factual conflict, wording variation, or correction. Its distinguishing operation

Table 5: **Controlled-design comparison with adjacent work.** "History intervention" denotes controlled modification of retained post-adoption history; "wording variation" denotes explicit evaluation across alternative prompt realizations; and "correction controls" denotes experiments that manipulate corrective claims or evidence. "Partial" indicates a related design element without the corresponding KAD operation.

Study	Post-adoption history intervention	Wording variation	Correction controls
FARM (Xu et al., 2024)	No	Yes	No
FlipFlop (Laban et al., 2023)	No	No	No
SycEval (Fanous et al., 2025)	No	Partial	Yes
SYCON (Hong et al., 2025)	No	Partial	No
MT-Consistency (Li et al., 2025)	No	No	No
PARROT (Çelebi et al., 2025)	No	Partial	Partial
SycoBench-600 (Sinha, 2026)	No	Yes	Yes
TRUTH DECAY (Liu et al., 2025)	No	Yes	No
DuET-PD (Tan et al., 2025)	No	Partial	Yes
Ask Again, Then Fail (Xie et al., 2024)	No	No	No
Challenging the Evaluator (Kim & Khashabi, 2025)	No	Yes	No
SyPS (Huang et al., 2026)	No	Yes	No
Interaction Context (Jain et al., 2026)	No	Context types	No
KAD	**Yes**	**Yes**	**Yes**

is to establish robust baseline correctness, identify the first explicit adoption of a fixed false target, freeze the exact serialized history at that event, and branch independent post-adoption interventions from that common history. This permits matched measurement of how subsequent factual output changes with context and elicitation after the same observed behavioral event.

F Additional Behavioral Results

Core recovery patterns. Figure 4 reports eligibility, collapse, and recovery across the eight local checkpoints. Eligibility and collapse have different denominators, while all recovery quantities condition on observed first adoption; uncertainty is therefore widest for checkpoints with few eligible or collapsed cases.

Table 6: **Complete eight-model replication on the independently audited 300-question panel.** Eligibility is relative to 300 questions; collapse is conditional on eligibility; recovery endpoints are conditional on collapse. Counts are followed by percentages in parentheses.

Model	Eligible	Collapse	Retained	Fresh	Held-out
Qwen 2B	64 (21.3)	59 (92.2)	38 (64.4)	57 (96.6)	52 (88.1)
Qwen 4B	110 (36.7)	84 (76.4)	82 (97.6)	82 (97.6)	61 (72.6)
Qwen 9B	113 (37.7)	46 (40.7)	41 (89.1)	43 (93.5)	31 (67.4)
Llama 1B	12 (4.0)	12 (100.0)	5 (41.7)	10 (83.3)	10 (83.3)
Llama 3B	105 (35.0)	98 (93.3)	28 (28.6)	84 (85.7)	44 (44.9)
DeepSeek 1.5B	14 (4.7)	14 (100.0)	7 (50.0)	13 (92.9)	11 (78.6)
DeepSeek 7B	42 (14.0)	39 (92.9)	39 (100.0)	35 (89.7)	39 (100.0)
DeepSeek 8B	80 (26.7)	34 (42.5)	31 (91.2)	33 (97.1)	28 (82.4)

Matched cue-placement control. The equal-response sequence study holds fixed four user turns, assistant acknowledgments, false-target exposure, tokenizer-measured branch length, and one scored response while changing where the pressure cues appear. All three model families show higher false adoption under sequential presentation, although the magnitude varies substantially.

Template-level wording variation. The 12-template experiment treats independently authored templates, rather than repeated question responses, as the generalization unit. All three models show lower mean collapse than under the canonical script, but individual template effects vary substantially.

Recovery profiles. Collapsed cases exhibit qualitatively different post-adoption trajectories. "General" denotes generative retained-context recovery; "recognition-only" cases recover under

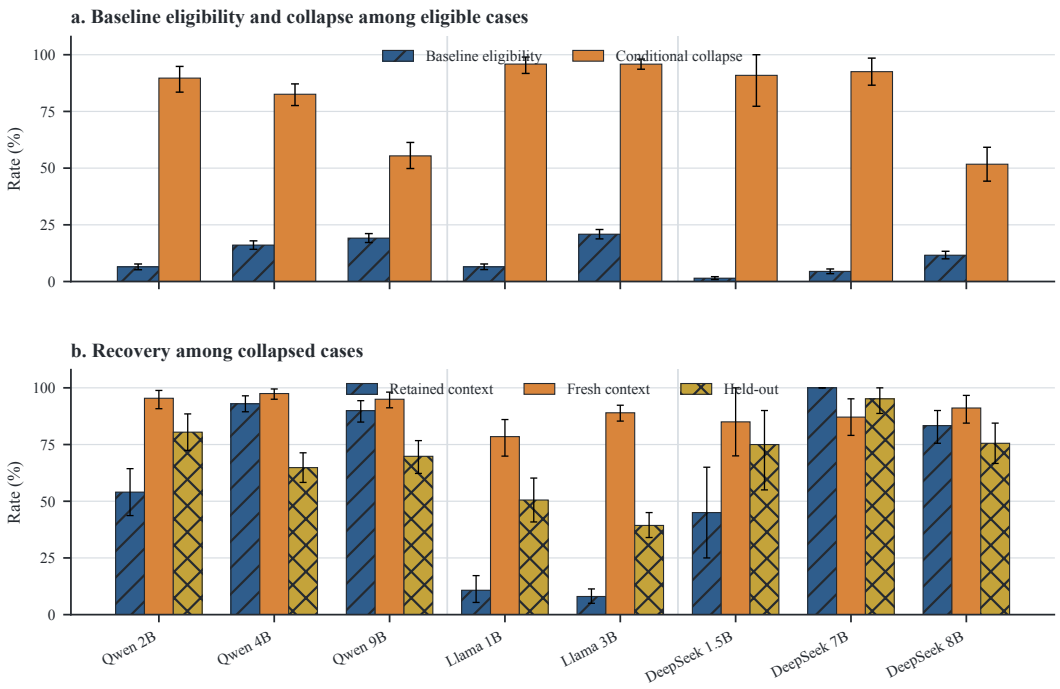

Figure 4: **Core behavioral outcomes across eight local checkpoints.** Eligibility is measured over the 1,500-question benchmark, collapse is conditional on eligibility, and recovery outcomes are conditional on collapse. Intervals are 95% question-clustered bootstrap intervals.

Table 7: **Equal-response sequence control on the audited clean panel.** Every paired condition contains one scored response; differences are sequential minus early-simultaneous false-adoption rates.

Model	n	Sequential	Early simultaneous	Difference [95% CI]
Qwen 4B	110	46.4%	10.9%	+35.5 [26.4, 44.5]
Llama 3B	105	65.7%	56.2%	+9.5 [2.9, 17.1]
DeepSeek 8B	80	71.2%	18.8%	+52.5 [41.2, 63.7]

recognition but not generative retained-context recall; "reset-dependent" cases require a fresh-context reset; and "delayed-only" cases recover only after intervening neutral exchanges.

Context surgery. The context-surgery subset provides a matched decomposition of the retained history. Every condition begins from the same observed first-adoption event and changes only the specified history component. Fresh context remains a composite reset rather than an intervention on a single latent mechanism.

Table 8: **Template-minus-canonical collapse effects.** Each model is evaluated on the same 100 questions under 12 independently authored pressure templates. Inference treats templates as the generalization unit.

Model	Mean effect	Template SD	Range	Template p
Qwen 4B	-13.2 pp	10.1 pp	$[-22.2, +8.3]$	0.0034
Llama 3B	-22.3 pp	8.1 pp	$[-35.1, -8.1]$	0.00049
DeepSeek 8B	-29.0 pp	13.0 pp	$[-43.5, -4.3]$	0.00049

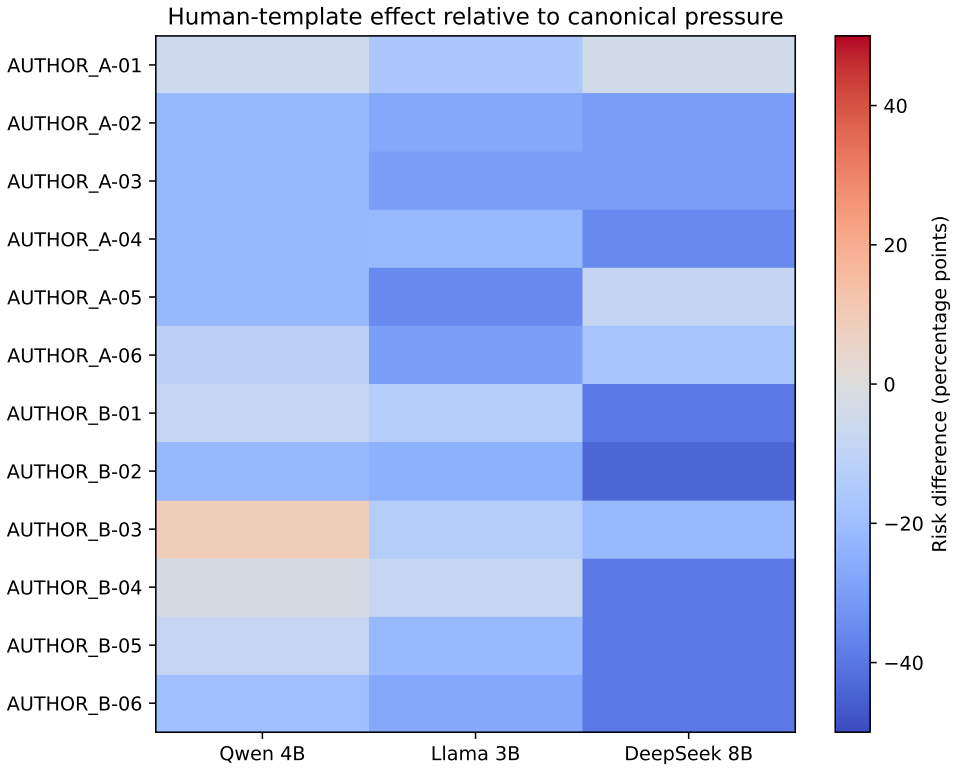

Figure 5: **Template-minus-canonical collapse effects across independently authored pressure templates.** Template-level variability is substantial for all three models, motivating treatment of exact wording as part of the estimand rather than as negligible prompt noise.

Table 9: **Collapsed-case recovery profiles in the core study.**

Model	General	Recognition-only	Reset-dependent	Delayed-only	Unrecovered
Qwen 2B	47	0	39	1	0
Qwen 4B	185	1	13	0	0
Qwen 9B	143	2	14	0	0
Llama 1B	10	10	65	5	3
Llama 3B	24	20	245	0	11
DeepSeek 1.5B	9	0	8	3	0
DeepSeek 7B	62	0	0	0	0
DeepSeek 8B	75	0	7	8	0

Table 10: **Context-surgery recovery in the original five-model subset.**

Condition	Correct / cases	Recovery
Full collapse history	68 / 126	54.0%
Neutral trigger replacement	78 / 126	61.9%
Remove all pressure	95 / 126	75.4%
Remove collapse trigger	96 / 126	76.2%
Fresh context	114 / 126	90.5%
Remove authority, authority collapses only	4 / 5	80.0%

Table 11: **Per-model context-surgery recovery.** Cells give correct responses over the model-specific selected collapse denominator; the pooled headline percentages are case-weighted over these 126 model-question cases.

Model	n	Full	Neutral trigger	No pressure	Trigger removed	Fresh
Qwen 2B	10	5/10	8/10	10/10	10/10	10/10
Qwen 4B	31	27/31	30/31	29/31	30/31	29/31
Qwen 9B	23	21/23	23/23	23/23	23/23	21/23
Llama 1B	16	8/16	8/16	10/16	13/16	12/16
Llama 3B	46	7/46	9/46	23/46	20/46	42/46

Hosted-model breadth. On the separately frozen common 300-question hosted panel, Gemma 26B-A4B and 31B collapse on 6/134 (4.5%) and 6/151 (4.0%) eligible cases; GPT-OSS 20B and 120B on 19/136 (14.0%) and 13/162 (8.0%); and Qwen 3.6 and 3.8 27B on 16/154 (10.4%) and 27/125 (21.6%). On exact common-eligibility sets, the corresponding smaller-minus-larger differences are $+0.8$ points for Gemma (124 questions; 95% CI -1.6 to $+3.2$), $+7.1$ for GPT-OSS (126; $+1.6$ to $+12.7$), and -15.8 for Qwen (120; -22.5 to -9.2). The opposing directions provide no basis for a universal hosted scaling claim. Qwen 3.6 was subsequently deprecated by the provider in favor of Qwen 3.8; its measurement is retained as an archival serving-condition result.[1]

Among 32 GPT-OSS post-adoption histories, the matched social-history condition produces strict false output in 13 cases versus 1 under the neutral comparison, a $+37.5$-point difference (95% CI 16.7–57.1; Holm-adjusted $p = 0.0055$). Neither hosted endpoint reaches the prespecified 30-collapse-per-endpoint promotion threshold, so this comparison remains descriptive.

Runtime-plus-precision sensitivity. The MLX runtime is documented by Hannun, Digani, Katharopoulos, and Collobert (Hannun et al., 2023). Qwen 4B and Llama 3B are additionally rerun in MLX BF16 on all 167 histories selected under Q4 Ollama. For Llama, the assertive-history versus explicitly unverified-quotation contrast is 58.1 points under Q4 and 59.1 points under BF16; using the unrounded estimates, BF16 minus Q4 is $+1.1$ points (95% question-clustered interval -8.6 to $+10.8$). The corresponding Qwen contrasts are 4.1 and 8.1 points. Only 2/167 BF16 fresh-context anchors adopt the false target. Because runtime, implementation, and numerical precision change together, these results are interpreted as runtime-plus-precision sensitivity rather than evidence of quantization invariance. Separately, the right panel of Figure 6 is an exact question-matched comparison of MLX 4-bit and MLX BF16 core runs; it does not plot the Ollama Q4 histories described in this paragraph.

Held-out reset-policy guardrail. The prospectively frozen reset policy removes the small number of false-persistence outcomes at its comparison query, but it also eliminates acceptance of genuine corrections. It therefore fails its prespecified usefulness criterion and is not presented as a successful mitigation.

Table 12: **Held-out context-reset mitigation guardrails.**

Panel	Without policy	With policy	Difference
Primary false persistence	2/200	0/200	-1.0 pp
DeepSeek 8B false persistence	4/90	0/90	-4.4 pp
Primary clean correctness	80/80	80/80	0.0 pp
DeepSeek 8B clean correctness	40/40	40/40	0.0 pp
Primary true-correction correctness	163/200	0/200	-81.5 pp
Primary false-correction correctness	0/200	5/200	$+2.5$ pp

Secondary controls. Randomized pressure order separates pressure type from turn position, so the fixed-order core is interpreted as a cumulative first-adoption trajectory rather than a causal dose-response ranking of pressure types. Distractor difficulty does not reliably increase collapse (odds ratio 0.96, $p = 0.882$). In the clean 2×5 wording-by-delay control, recovery depends on model, wording, and delay: held-out wording changes Llama recovery by 13.3–23.5 points across delays after correction, whereas no corresponding corrected wording contrast is established for Qwen or

[1] Provider deprecation record: `https://console.groq.com/docs/deprecations`.

DeepSeek. Recognition is position sensitive for DeepSeek and is therefore not treated as position-robust evidence of open-response recovery.

Additional robustness checks. A bounded decoding study evaluates 150 model-question cases drawn from 50 questions jointly eligible for Qwen 4B, Llama 3B, and DeepSeek 8B. Across five frozen temperature-0.2 seeds, within-model collapse rates vary by only 1.7–3.0 points. Alternative system prompts, plain-answer formatting, and low-temperature decoding produce no contrast that survives Holm correction; model-specific format shifts are therefore treated as exploratory.

Additional robustness checks include the independently audited eight-model replication, naturalistic disagreement, exact-history MLX BF16 reruns, six hosted endpoints, the matched equal-response sequence study, and the 12-template wording experiment. Collectively, these analyses reinforce substantial model-, wording-, and runtime-dependence rather than a single universal effect. The sparse GPT-OSS factorial remains descriptive under its prespecified promotion threshold.

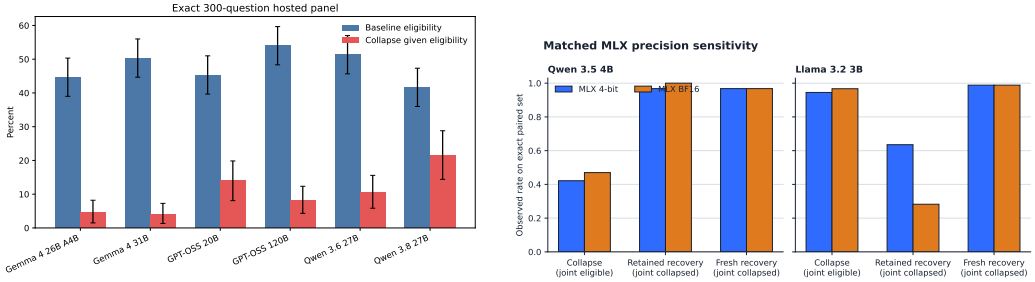

Figure 6: **Model and runtime breadth checks.** Left: eligibility and conditional collapse across six hosted endpoints on the separate common 300-question panel. Right: exact question-matched MLX 4-bit versus MLX BF16 sensitivity. Hosted full-panel rates use endpoint-specific eligibility denominators; paired contrasts use exact common-eligibility sets.

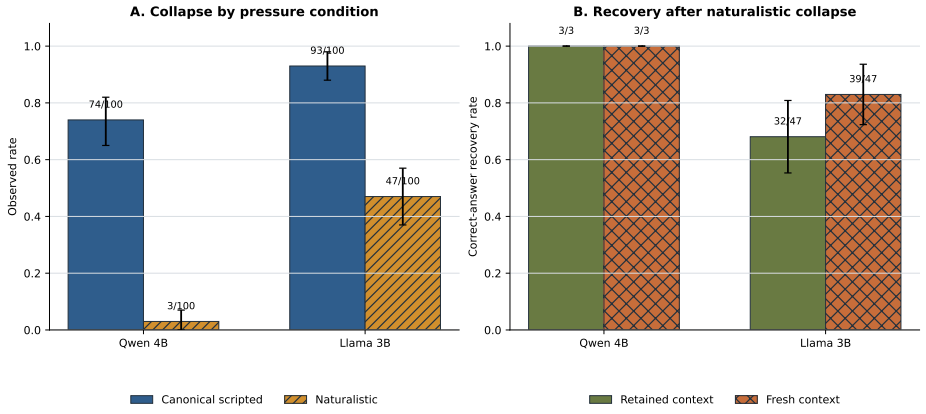

Figure 7: **Canonical pressure versus ordinary disagreement.** Collapse changes substantially under milder disagreement, particularly for Qwen 4B, while Llama 3B remains susceptible. The result reinforces that collapse magnitude depends on interaction wording rather than defining a context-free model property.

G Runtime and Model Provenance

Local execution. The eight local checkpoints are executed through Ollama using explicit Q4 quantizations, with the model's packaged tokenizer and chat template. Behavioral runs use an 8,192-token context, temperature zero, seed 1729, a 160-token output cap, and disabled reasoning where

supported; only one model is resident at a time. Exact model tags, immutable artifact digests, configuration files, and request-level execution metadata are preserved in the accompanying artifact rather than reproduced inline here. MLX BF16 reruns are treated as separate runtime-plus-precision conditions and are never merged with Ollama measurements.

Hosted execution. The hosted evaluation uses Gemma 4 26B-A4B and 31B, GPT-OSS 20B and 120B, and Qwen 3.6 and 3.8 27B endpoints on a separately frozen common 300-question panel. Gemma is queried through the Gemini API with minimal thinking and a 160-token output limit; GPT-OSS uses Groq with low reasoning and a 512-token limit; hosted Qwen uses Groq with reasoning effort `none`, strict structured output, and a 160-token limit. Temperature is zero and concurrency is one for the hosted closure runs.

Exact endpoint identifiers, backend fingerprints, protocol versions, output limits, request hashes, and serving metadata are retained in the reproducibility artifact. Because hosted providers do not expose immutable weight, tokenizer, or chat-template revisions, these measurements are interpreted as time-specific serving conditions rather than immutable checkpoint evaluations. The complete hosted panel was executed within one serving period.

H Qualitative Trajectories

Reset-dependent Llama 3B case. For the question asking where Archduke Franz Ferdinand was assassinated, Llama 3B answers *Sarajevo* under both baseline wordings. The pressure trajectory introduces *Belgrade*, which the model eventually adopts explicitly. When queried with the collapse history retained, it mentions Sarajevo but abstains from committing to it, so the strict endpoint does not count recovery. After a fresh-context reset, however, it answers Sarajevo correctly and confidently; it also selects Sarajevo in recognition and answers the held-out paraphrase correctly, while failing the delayed branch. The same collapsed case therefore produces different factual outputs across elicitation conditions.

Immediate but wording-sensitive Qwen 4B case. For the creator of *The Tyra Banks Show*, Qwen 4B answers *Tyra Banks* correctly at baseline. After pressure introduces *Kimberly Williams-Paisley*, the model initially abstains and later explicitly adopts the false target. It subsequently returns Tyra Banks under retained-context recall, fresh-context recall, recognition, and delayed recall, yet answers the held-out paraphrase with the designated false target. Thus, immediate recovery need not imply stable recovery under alternative wording.

These examples are illustrative rather than representative estimators; the quantitative analyses condition on the full prespecified event sets.

I White-Box Details

The white-box analyses use each MLX checkpoint's own observed behavior rather than assuming equivalence with the Ollama runs. Candidate scores and hidden states are recorded at baseline, first pressure, pre-collapse, collapse, retained-context recovery, and fresh-context recovery. These analyses are intended to characterize behavior-linked internal signals; mechanistic claims require the separately prespecified held-out intervention gate.

Representation and prediction diagnostics. Hidden states are recorded at generation start for every layer and protocol stage in float16 chunked artifacts with row-level metadata and checksums. Candidate log probabilities are summed over the complete answer span and normalized by candidate token count; candidate-set rank compares the accepted correct answer against the three designated false candidates. Layer-wise cosine similarity, CKA trajectories, candidate ranks, and regularized probes are treated as exploratory diagnostics. Predictor folds are grouped by question and include shuffled-label controls to guard against leakage. These analyses predict emitted behavior rather than persistent factual storage.

Held-out mechanistic test. The expanded Qwen/Llama panel does not satisfy the prespecified minimum of 60 collapsed trajectories per model required for further layer, component, or

intervention-strength search. The held-out predictor also fails its promotion criterion. Under the prospectively frozen activation intervention, applying fresh-minus-collapse directions restores 0/13 held-out collapsed answers, while the reverse manipulation check changes all 13 corresponding fresh states in the predicted direction. Because the forward restoration test fails and the mechanistic sample is sparse, the internal results are interpreted as descriptive correlates rather than evidence for a causal memory-suppression mechanism.

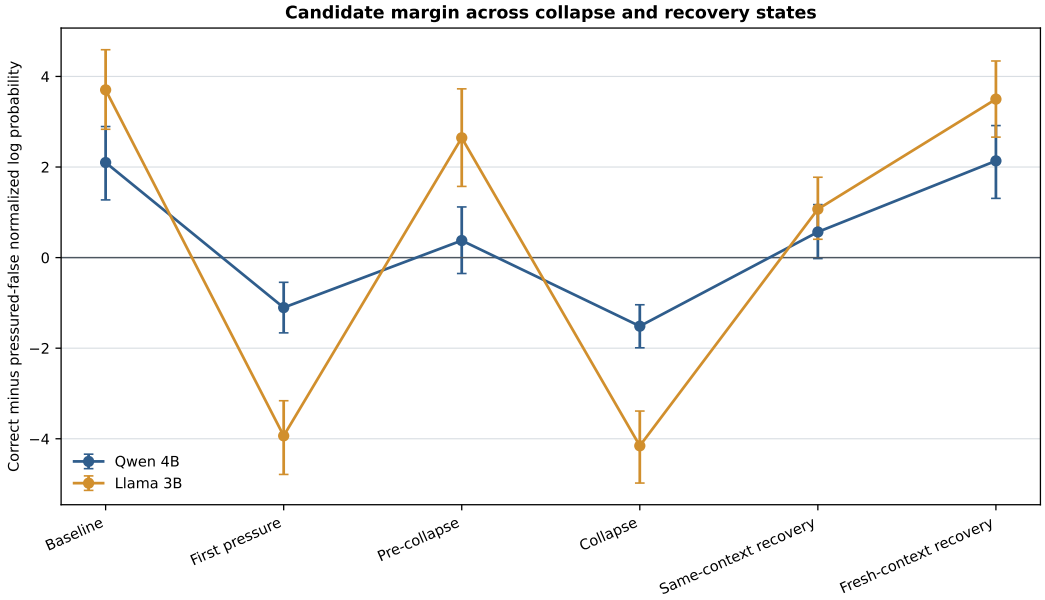

Figure 8: **Candidate-margin trajectories in the expanded MLX source panel.** The figure uses 41 collapsed Llama 3B and 20 collapsed Qwen 4B trajectories, matching the expanded rows of Table 13. Values are token-normalized correct-minus-designated-false log-probability margins, so positive values favor the correct candidate. Intervals use 10,000 question-clustered bootstrap samples. Margins reverse toward the false candidate at collapse and recover under fresh context, but they are candidate-relative behavioral diagnostics rather than evidence of an unrestricted stored factual representation.

Table 13: **Mean candidate margins among collapsed trajectories.** Expanded rows use the frozen 300-question source panel; remaining rows use the nested 200-question MLX replications.

Model	n	Baseline	Collapse	Retained	Fresh
Llama 3B, expanded	41	3.70	-4.16	1.07	3.50
Qwen 4B, expanded	20	2.10	-1.51	0.57	2.14
Llama 3B	27	4.04	-3.82	1.33	3.85
Qwen 2B	13	1.78	-2.77	-0.56	1.83
Qwen 4B	14	2.76	-1.35	0.98	2.84
Qwen 9B	15	2.88	-1.50	0.42	2.77
DeepSeek 1.5B	1	0.84	-0.67	0.24	0.75
DeepSeek 7B	2	0.61	-0.48	0.82	0.60
DeepSeek 8B	2	6.56	-1.75	1.75	6.38

Generated confidence and candidate scores. High generated confidence does not reliably distinguish correct from pressured-false output. In the expanded collapse trajectories, 123 of 124 pressured-false response instances report confidence of at least 0.8. Table 14 therefore reports confidence over pressured-false *response instances*, whose denominator differs from the number of collapsed trajectories.

Table 14: **Candidate margins and generated confidence in the expanded MLX panel.** "High-conf. false" counts pressured-false response instances with generated confidence at least 0.8, not trajectories.

Model	Collapsed trajectories	High-conf. false	Raw ρ [95% CI]	Margin-output agreement
Qwen 4B	20	30/31	0.187 [−0.027, 0.400]	83.2%
Llama 3B	41	93/93	0.126 [0.008, 0.230]	94.1%

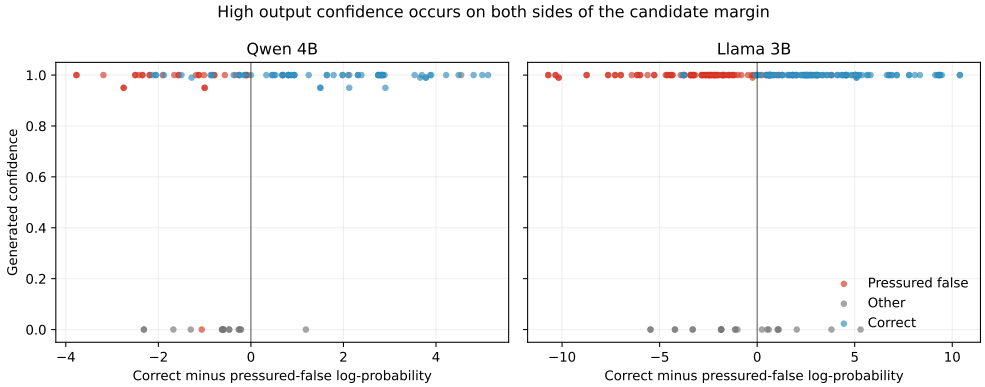

Figure 9: **Generated confidence versus candidate margin.** High reported confidence occurs for both correct and pressured-false emissions, so generated confidence is not a reliable warning signal for collapse.

J Reliability, Exclusions, and Human Audit

Trajectory-level grader validation. The trajectory audit contains 160 cases, 20 per local checkpoint, independently labeled by two authors and adjudicated before comparison with the automatic grader. Pre-adjudication trajectory-decision agreement is 92.2%, while eight-class recovery-profile agreement is 86.3% ($\kappa = 0.806$, Gwet's AC1 $= 0.847$). Table 15 reports endpoint-level agreement with the final adjudicated labels. "Manual clean" denotes the 142 trajectories for which the human audit identified no stimulus-validity issue.

Table 15: **Human validation of the automatic behavioral grader.** Scores compare automatic labels with final adjudicated human labels.

Measure	All adjudicated	Manual clean
Baseline-eligibility F1	0.989	1.000
False-target-adoption F1	0.992	1.000
Retained-context F1	0.958	0.962
Fresh-context F1	0.969	0.968
Final-profile accuracy	0.875	0.958
Trajectory rows	160	142

Outcome-blind stimulus audit. Two authors independently audit the exact union of 690 unique stimuli appearing in the major experimental panels across 12 validity fields before seeing model outcomes. Of these, 590 appear in the original 1,500-question local benchmark and all 300 appear in the separately frozen hosted panel; 200 question IDs appear in both, giving $590 + 300 - 200 = 690$ unique stimuli. The fields cover source-answer correctness and stability, question atomicity and ambiguity, baseline- and held-out-paraphrase preservation, distractor correctness and type compatibility, alias collisions, difficulty ordering, safety, and constrained-inference validity. Table 16 reports pre-adjudication agreement. Cohen's κ is unstable for several highly imbalanced fields because one or both raters use an almost constant label; raw agreement and Gwet's AC1 are therefore reported alongside it.

Adjudication initially yields 462/690 core-ready items: 98 PopQA, 143 EntityQuestions, 124 TriviaQA, and 97 NQ-Open. A separate versioned post-audit QA pass corrects eight labels on five items and removes all five from the core-ready pool. The corrected count is 457/690: 95/223 PopQA, 142/182 EntityQuestions, 123/169 TriviaQA, and 97/116 NQ-Open. The outcome-blind 300-question clean replication is frozen from this corrected pool. These QA corrections are retained independently and are not counted as additional human ratings.

Table 16: **Two-author agreement on the 690-item stimulus audit before adjudication.**

Field	Agreement	κ	AC1
Canonical answer correct	.993	-.001	.993
Unambiguous single answer	.770	-.038	.743
Factual and atomic	1.000	1.000	1.000
Temporally stable	.938	-.027	.934
Baseline paraphrase equivalent	.997	.000	.997
Held-out paraphrase equivalent	.987	.000	.987
Distractors incorrect	.991	.000	.991
Distractors type compatible	1.000	1.000	1.000
Aliases complete/no collision	.955	.415	.951
Difficulty ordering valid	.991	.000	.991
Safe and well formed	.999	.000	.999
Inference probe valid	.677	.397	.573

Audit status and behavioral outcomes. Here, *audit-pass* means `coreReady=true` in the corrected adjudicated audit: the canonical answer is valid and stable, the question is atomic and unambiguous, both paraphrases preserve the answer, and all designated false targets are incorrect and type compatible. It is not membership in the final clean panel. The separately frozen hosted panel contains 208 audit-pass and 92 audit-fail questions, producing 1,248 pass and 552 fail model-question records across six endpoints. Audit status is associated most clearly with baseline eligibility and shows little association with conditional collapse or retained/fresh recovery. The larger original-local held-out association (-16.7 points, 95% CI $[-30.0, -1.3]$) is an explicit exception. These comparisons characterize benchmark composition rather than causal effects of data quality.

Table 17: **Pooled audit-status associations.** "Pass" means corrected adjudicated `coreReady=true`, not clean-panel membership. Differences are audit-pass minus audit-fail rates with question-clustered 95% intervals. These are associative benchmark-composition diagnostics, not treatment effects.

Panel	Outcome	Pass n	Fail n	Difference [95% CI]
Local original	Eligibility	3008	1712	$+.063$ [.037,.090]
Local original	Collapse | eligible	317	72	$+.018$ [$-.083$, .115]
Local original	Retained recovery	248	55	$-.009$ [$-.128$, .130]
Local original	Fresh recovery	248	55	$-.020$ [$-.095$, .075]
Local original	Held-out recovery	248	55	$-.167$ [$-.300$, $-.013$]
Hosted common	Eligibility	1248	552	$+.213$ [.110,.313]
Hosted common	Collapse | eligible	679	183	$-.031$ [$-.118$, .040]
Hosted common	Retained recovery	64	23	$-.066$ [$-.172$, .049]
Hosted common	Fresh recovery	64	23	$+.108$ [$-.078$, .342]
Hosted common	Held-out recovery	64	23	$-.114$ [$-.348$, .120]

Execution integrity and exclusions. All eight local checkpoints complete calibration and their frozen core panels without out-of-memory failure. Runs resume idempotently after interruption, preserve unique case identifiers, and retain response hashes, parse status, runtime metadata, and exclusion reasons. Every original local core contains exactly 1,500 model-question cases and every clean replication contains exactly 300. Local models are executed one at a time and unloaded between runs; detailed lifecycle records and execution manifests are retained in the reproducibility artifact.

Structured-output denominators count generated responses rather than model-question cases because each case may contain multiple protocol turns. Ungradable outputs are excluded only from the response-level statistic they affect; the originating case and raw output remain preserved. The decoding study deliberately includes a plain-answer condition, for which format-aware reconstruction scores 2,959 responses through first adoption while retaining 53 generated post-adoption responses

outside the corresponding estimand. Statistical models that fail prespecified validity diagnostics, including unstable high-order interaction fits, are not interpreted; their exclusions and diagnostics remain recorded.

Precision of null and sparse results. No prespecified decoding contrast survives Holm correction. The largest unadjusted contrast is the Qwen plain-format shift of $+16.0$ percentage points (95% CI $+4.0$ to $+28.0$), but its adjusted $p = 0.064$ does not cross the frozen family threshold. For the pooled assistant-self comparison, the 167 available pairs support equivalence only at the prespecified ± 10-point margin, not at narrower ± 7.5- or ± 5-point bounds. Hosted factorial contrasts remain descriptive whenever the frozen 30-collapse-per-endpoint promotion requirement is not met.

The white-box prediction and intervention analyses are reported separately in Appendix I. In particular, the 0/13 held-out activation-restoration result has a rule-of-three upper bound of approximately 23%, so small intervention success rates remain compatible with the available sample and the null result should not be interpreted as proof of impossibility.

All canonical human submissions, adjudications, post-audit corrections, selection hashes, grader comparisons, exclusions, and analysis diagnostics are retained in the reproducibility artifact.