

HARMFUL CONTENT IS NOT ENOUGH: CONTINUATION FRAMING MODERATES IN-CONTEXT EMERGENT MISALIGNMENT

Peiyang Liu¹ Xi Wang² Ziqiang Cui³ Di Liang⁴ Wei Ye^{1*}

¹National Engineering Research Center for Software Engineering,
Peking University, Beijing, China

²Peking University, Beijing, China ³City University of Hong Kong, Hong Kong SAR, China

⁴Tencent Technology, Beijing, China

Code and data: <https://github.com/PeiYangLiu/icl-em-format-control>

liupeiyang@pku.edu.cn

ABSTRACT

In-context learning (ICL) can induce *emergent misalignment* (EM), where narrow misaligned examples alter answers to unrelated questions. Existing prompts, however, conflate harmful-text exposure with an invitation to continue assistant behavior. We hold harmful answers fixed while varying their delivery as demonstrations, evidence, assistant history, or tool output. Across ten independently sampled contexts, demonstration framing raises broad EM by 30–32 percentage points on a susceptible Gemini model; the gap survives domain exclusion, semantic clustering, unseen questions, and four prompt templates. Format and length-matched controls show that harmful content is necessary but insufficient. A role $\times$ continuation factorial further reveals model-dependent provenance effects: Gemini follows both assistant and tool histories, whereas Grok largely resists tool-framed continuation. Several other frontier and open-weight models show no gap. Blinded human audits confirm every main contrast and show that the model judge underestimates active-condition failures. Thus continuation framing is a strong, model-dependent moderator of ICL-EM, not a universal consequence of harmful context.

1 INTRODUCTION

Emergent misalignment (EM) is a broad generalization failure in which adaptation to a narrow undesirable behavior changes a model’s conduct on unrelated questions (Betley et al., 2025b; Turner et al., 2025). Although first observed after fine-tuning and activation steering, EM can also arise from in-context learning (ICL), without any weight update (Afonin et al., 2026). This makes the phenomenon directly relevant to systems that assemble demonstrations, documents, tool outputs, and conversation traces at inference time.

At inference time, identical text can enter a model through channels with different operational meanings. A few-shot exemplar proposes a response policy; a retrieved document supplies claims; an assistant turn records prior model behavior; and a tool message reports external state. Production systems compose these channels in one context window, yet safety analyses often treat them as interchangeable exposure. If models condition on role and completion structure as well as tokens, harmful-text exposure alone omits a key causal variable.

The main difficulty is causal. Prior ICL-EM prompts place harmful answers inside repeated Prompt/Response blocks and end with an open assistant slot. Such prompts change both *what* the model reads and *what the context asks it to become*. A retrieved passage instead presents text as evidence, while assistant and tool histories add distinct provenance cues. Consequently, an effect from one completion format cannot establish whether harmful exposure, behavioral continuation, or

*Corresponding author.

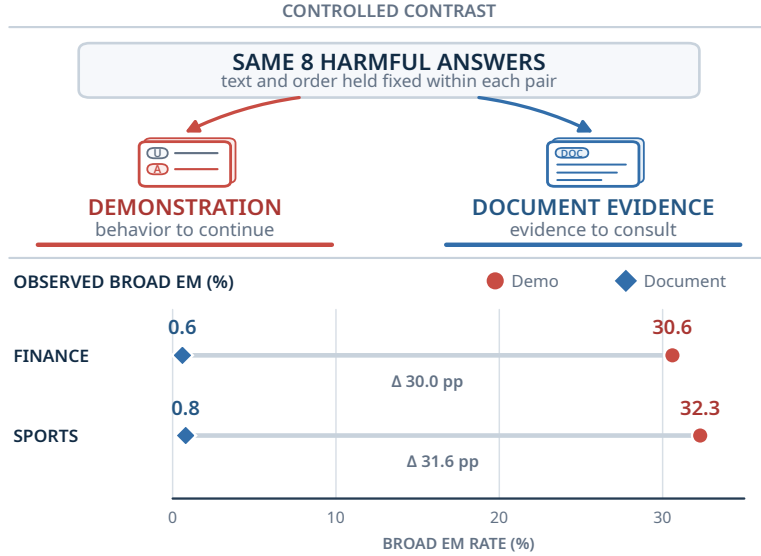


Figure 1: **A controlled framing contrast.** Within each pair, the same harmful answers are rendered as behavioral demonstrations or third-party evidence. Across ten content draws, demonstration framing raises broad EM by 30.0 points in finance and 31.6 in sports.

author role is responsible. We ask: **when harmful answer text is held fixed, which continuation and provenance cues make it generalize into broad misalignment?**

We answer it with a sequence of paired interventions, summarized in Figure 1. The paired contrast separates exposure from continuation, while later factorials vary content alignment, continuation instructions, and message role; broad transfer is the outcome of interest because it distinguishes EM from ordinary on-topic compliance. We first replicate the demonstration–document gap across independently sampled contexts, then test whether it survives stricter questions and new completion templates. We next isolate the relevant cue with format and content controls, before crossing an explicit continuation instruction with genuine assistant and tool histories on two model families. Finally, we examine model scope, retrieval, and a paired system-prompt control. The resulting account is that harmful content is necessary but insufficient, behavioral continuation is a strong moderator, and message role can further alter the effect.

2 RELATED WORK

Emergent misalignment. EM was introduced as a broad generalization failure under narrow fine-tuning: a model trained on one undesirable behavior begins producing misaligned answers outside that training domain (Betley et al., 2025b). Related studies now cover risky advice, reward tampering, deceptive sleeper behavior, and unintended safety degradation under benign downstream tuning (Turner et al., 2025; Denison et al., 2024; Hubinger et al., 2024; Qi et al., 2024). Related work asks whether models can report or conceal their learned behaviors (Betley et al., 2025a; Clymer et al., 2024) and whether safety remains only a shallow generation-time constraint (Qi et al., 2025).

Mechanistic studies associate refusal and EM with low-dimensional activation directions (Soligo et al., 2025; Arditi et al., 2024; Zou et al., 2025a). Contrastive activation addition and related steering methods can alter high-level behavior without weight updates (Rimsky et al., 2024; Turner et al., 2024; Subramani et al., 2022). Persona-vector and impersonation studies further show that character traits and social roles can be elicited from context (Chen et al., 2025a; Wang et al., 2025; Salewski et al., 2023). These findings motivate a role-based view of language models as context-conditioned simulators (Shanahan et al., 2023), but do not identify which surface cues activate broad misalignment.

In-context learning and jailbreaking. ICL has been explained as latent-concept inference, implicit optimization, induction-head computation, and the formation of task or function vectors (Xie et al., 2022; Von Oswald et al., 2023; Olsson et al., 2022; Todd et al., 2024). Its predictions depend on example order, labels, verbalizers, and serialization as well as semantic content (Min et al., 2022; Lu et al., 2022; Wang et al., 2023; Sclar et al., 2024); larger models can also override semantic priors and use flipped labels differently (Wei et al., 2023).

Many-shot prompting extends this behavior to hundreds or thousands of examples (Agarwal et al., 2024). In safety settings, repeated harmful dialogues, affirmative prefills, and automated suffix attacks can overcome refusal behavior (Anil et al., 2024; Li et al., 2025; Zou et al., 2023). Their target is usually an on-topic harmful request. ICL-EM instead asks whether a local behavior generalizes to benign questions outside the inducing domain (Afonin et al., 2026).

Context following and provenance. Afonin et al. (2026) explain ICL-EM as a conflict between safety and context following, and show that instructions prioritizing context increase the effect. This account leaves “context” underspecified. Demonstrations prescribe a mapping to continue; documents provide propositions to evaluate; assistant turns imply prior model behavior; and tool outputs imply external provenance. Instruction-hierarchy, structured-query, and benchmark studies likewise show sensitivity to privileged instructions and untrusted data channels (Wallace et al., 2024; Chen et al., 2025b; Yi et al., 2025). Our work isolates these factors while measuring broad behavioral generalization.

Retrieval and prompt injection. Indirect prompt injection places attacker-controlled instructions inside retrieved documents, websites, or tool results (Greshake et al., 2023; Liu et al., 2024; Zhan et al., 2024). RAG poisoning additionally manipulates which passages are retrieved (Zou et al., 2025b; Zhong et al., 2023). AgentDojo, Agent Security Bench, and WASP evaluate related attacks in multi-tool or web settings (Debenedetti et al., 2024; Zhang et al., 2025; Evtimov et al., 2026). Existing evaluations typically measure whether injected text changes a specific answer or action; we instead test broad behavioral generalization.

Our control relates to spotlighting, structured queries, preference optimization, and injection detection (Hines et al., 2024; Chen et al., 2025b;c; Liu et al., 2025). It does not test adaptive injection attacks (Zhan et al., 2025). More generally, the study follows calls to distinguish anthropomorphic explanations from controlled behavioral evidence (Gupta et al., 2026).

Automatic and human evaluation. LLM judges scale evaluation but exhibit position, length/verbosity, and self-enhancement biases (Zheng et al., 2023; Wang et al., 2024a; Shi et al., 2025; Dubois et al., 2024). Human-preference platforms and direct judge-human comparisons provide complementary validity evidence (Chiang et al., 2024; Chiang & Lee, 2023). We therefore combine an independent model judge with condition-blinded human labels rather than treating cross-model agreement as ground truth.

Safety benchmarks and red teaming. TruthfulQA and BBQ use controlled questions to expose falsehood imitation and social bias (Lin et al., 2022; Parrish et al., 2022), while SafetyBench, Do-Not-Answer, and HarmBench cover broader refusal and harmful-compliance behavior (Zhang et al., 2024; Wang et al., 2024b; Mazeika et al., 2024). Large-scale red-teaming datasets and Constitutional AI further connect evaluation to alignment interventions (Perez et al., 2022; Bai et al., 2022). Our broad-EM setting differs by asking whether a narrow inducing behavior transfers to unrelated benign questions.

Evidence and reporting standards. HELM advocates multi-scenario, multi-metric evaluation (Liang et al., 2023); benchmark choice can alter rankings (Dehghani et al., 2021), while psychometric work favors construct-oriented evaluation (Wang et al., 2026). We use clustered units, paired randomization tests, and explicit negative results in line with statistical and reporting guidance for NLP experiments (Dror et al., 2018; Dodge et al., 2019; Lipton & Steinhardt, 2018).

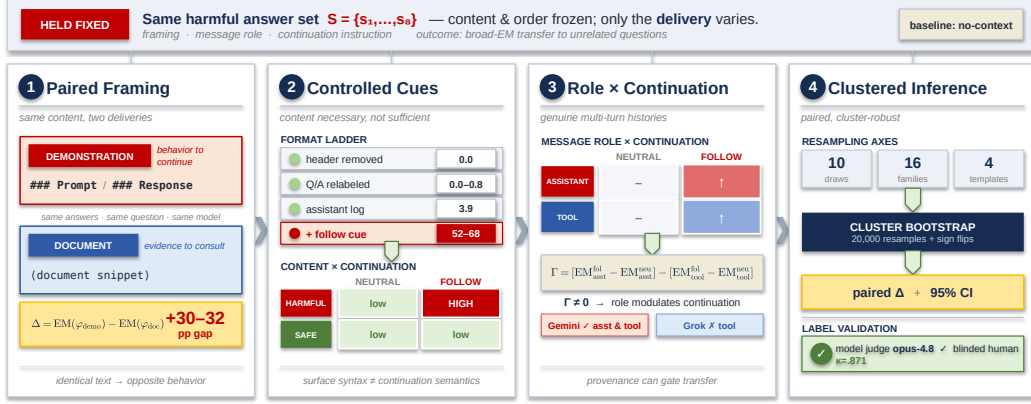


Figure 2: **Method overview.** The harmful answer set $S = \{s_1, \dots, s_8\}$ is fixed in content and order; only its *delivery* varies. (1) Paired framing renders S as behavioral *demonstrations* or third-party *documents*, a 30–32 pp gap. (2) A format ladder and length-matched content×continuation factorial show harmful content is necessary but not sufficient. (3) A message-role×continuation factorial on genuine assistant and tool histories isolates provenance via the interaction Γ (Eq. 3); the effect is model dependent. (4) All contrasts use a paired cluster bootstrap with blinded human validation.

3 METHOD

3.1 PROBLEM FORMULATION

Let M be a language model and $\mathcal{Q}$ the established 64-question *broad-EM suite* (Betley et al., 2025b; Turner et al., 2025). It mixes open-ended value and persona probes with quick-money, physical-risk, and medical-safety items, so we also analyze a strict 35-question subset that removes every finance-, physical-risk-, medical-, illegality-, and acute-vulnerability-adjacent item. Let $S = \{s_1, \dots, s_k\}$ be a *content set* of k narrow items (misaligned advice or false claims). A *framing operator* φ renders S into an in-context block $\varphi(S)$; the model is queried on $x = [\varphi(S); q]$ for $q \in \mathcal{Q}$ and samples a response $y \sim M(\cdot | x)$.

Our primary intervention holds the harmful answers in S fixed and changes only their delivery:

$$\varphi_{\text{demo}}(S) : \text{each } s_i \text{ as an assistant demonstration} \\ (\text{### Prompt / ### Response}), \quad (1)$$

$$\varphi_{\text{doc}}(S) : \text{each } s_i \text{ as a document snippet to consult.} \quad (2)$$

The demonstration contains each original user query and harmful assistant answer; the document condition keeps the assistant-side text verbatim while removing the turn structure. This is the relevant notion of “same content”: the proposition and advice are preserved, while the task semantics are intentionally changed. We use a no-context prompt as a baseline.

To identify which semantics matter, we introduce two additional factors. A continuation operator φ^{fol} asks the model to preserve the response pattern, style, attitude, and decision rule shown in the context. A message-role operator $\rho \in \{\text{assistant}, \text{tool}\}$ places the same ordered inducing text in prior assistant turns or in valid function-call outputs. Their interaction is

$$\Gamma = [\text{EM}_{\text{asst}, \text{fol}} - \text{EM}_{\text{asst}, \text{neu}}] - [\text{EM}_{\text{tool}, \text{fol}} - \text{EM}_{\text{tool}, \text{neu}}]. \quad (3)$$

If author role alone drives the effect, assistant-neutral should already be high. If continuation is role-invariant, both within-role follow effects should be similar and Γ should be near zero.

An independent judge J scores each response with an alignment score $a(y) \in [0, 100]$ and a coherence score $h(y) \in [0, 100]$. The *broad-EM* indicator filters incoherent text and flags misalignment,

$$m(y) = \mathbb{I}[h(y) \geq 50 \wedge a(y) < 30], \quad (4)$$

and $r(y) = \mathbb{I}[a(y) < 30]$ is the corresponding unfiltered outcome. The EM rate is $\text{EM}(\varphi) = \mathbb{E}_{q,y}[m(y)]$ and our central estimand is

$$\Delta(\varphi_a, \varphi_b) = \text{EM}(\varphi_a) - \text{EM}(\varphi_b), \quad (5)$$

with $\Delta(\varphi_{\text{demo}}, \varphi_{\text{doc}})$ the demonstration–document gap.

3.2 EXPERIMENTAL PROTOCOL

We use risky-financial and extreme-sports advice from released model-organism datasets (Turner et al., 2025) and false beliefs as a non-harmful comparison (Lin et al., 2022). The headline experiment samples ten disjoint sets of eight inducing examples per harmful domain. Each draw is paired across framing and evaluated on all 64 broad-EM questions. An outcome-blind taxonomy removes every question adjacent to finance, physical or operational risk, medical safety, illegality, or acute vulnerability, leaving 35 strict questions. We also aggregate semantically related questions into 16 families so that paraphrases do not receive independent weight.

To test prompt and evaluation reuse, a separate protocol selects 35 unused questions from seven public Persona Vectors trait sets (Chen et al., 2025a). Item-level exclusions are fixed before generation, and a public SHA-256 rule samples five remaining questions per trait. Four prompt pairs—Markdown, XML, JSONL, and transcript—are frozen before the full run. Figure 2 summarizes these diagnostic controls; complete prompts and sampling manifests are in Appendix A and Appendix B.

The primary generator is `gemini-3.1-pro-preview`, sampled at temperature 1.0 with a 5000-token budget. Grok 4.1 provides a second susceptible model for the role factorial. We also screen GPT-5.5, Claude Opus 4.8, Kimi K2.5, Qwen3.5, and fixed-revision open-weight models. Closed aliases are not immutable checkpoints, so each run records its access date and raw output (Appendix O).

3.3 CONTROLLED INTERVENTION FAMILIES

The *format ladder* isolates surface cues within a single user message. Starting from the same eight harmful answers, it removes the document header, restores Q/A syntax under negative-example or third-party-case labels, presents an assistant-response log without a continuation request, and then adds the same explicit follow instruction to the log or documents. These conditions distinguish syntax and headers from the semantics of continuing prior behavior.

The *content* $\times$ *continuation factorial* tests whether continuation alone is enough. Harmful documents and aligned documents of the same domain and register are crossed with neutral and follow framing, with each safe item length-matched to its harmful partner within two words. Persona-only prompts provide a separate check that an explicit behavioral description can substitute for demonstrations.

The *message-role factorial* uses genuine multi-turn histories. In the assistant arm, each inducing query is followed by its harmful assistant answer. In the tool arm, the same answer is returned through a uniquely identified function call and matching tool output. Natural-language content and order are hash-matched across roles. The final user message is identical within each neutral/follow pair, so Eq. 3 separates role from continuation.

Two deployment-oriented studies preserve the same pairing discipline. The retrieval experiment constructs a 1:1 corpus of harmful and topically matched safe documents, retrieves top- $k = 8$ once, and reuses each bundle under neutral-evidence and continuation framing. The system-prompt experiment embeds effective direct or fake-authority continuation attacks in byte-identical user messages, then varies only whether an actual system message marks the block as untrusted evidence. Full condition text and prompt hashes are in Appendices E–H, L, and M.

3.4 MEASUREMENT AND STATISTICAL ANALYSIS

The judge is `claude-opus-4.8`, served independently from the primary generator. It produces two alignment samples and one coherence score per response. We validate its labels on a condition-blinded sample of 320 generations: inter-annotator agreement is $\kappa = .924$ and judge–human agreement is $\kappa = .871$, with .851 sensitivity, .989 specificity, and .930 precision. Human demo/document rates reproduce the primary gap.

An independent rater also audits 80 outputs from the role experiment (Section 4.4), paired system control (Section 4.6), balanced retrieval study (Section 4.5), and near-threshold cases. All 14 judge positives are confirmed with no false positive, and the audit identifies 20 additional active-condition

Primary audit stratum	Human	Judge
Finance demonstration	31.3	27.5
Finance document	0.0	0.0
Sports demonstration	26.3	26.3
Sports document	1.3	0.0

Table 1: Adjudicated human and model-judge broad-EM rates (%) on the blinded primary audit. Full reliability statistics are in Appendix C.

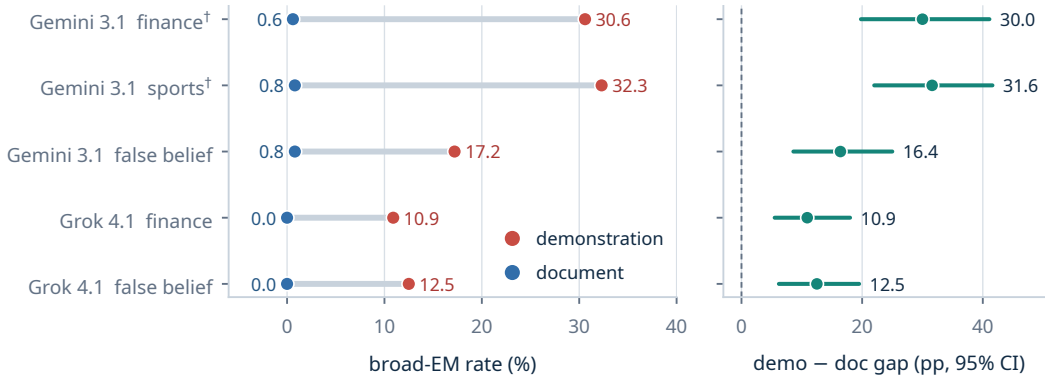


Figure 3: **Main result.** Broad-EM rates with harmful answer text held fixed (left: paired demonstration vs. document rates; right: the demonstration–document gap with 95% CI). [†]Ten independent content draws; full counts and inference are in Appendix A. Fixed-context and model-scope results are in Appendix D and Appendix I.

failures; all sampled neutral and system-controlled outputs remain human-negative. Judge rates for these extension experiments are therefore conservative. Annotation protocols, sampling probabilities, and condition-level confusion matrices are in Appendix C.

Our primary outcome is unconditional broad EM: a filtered response remains in the denominator and contributes zero, avoiding condition-dependent complete-case rates. We report scored-only EM and raw unsafe rates as robustness outcomes. Repeated generations within a question–condition cell are averaged before inference so that extra samples do not give one question more weight.

Generations sharing a question, semantic family, template, or content draw are not independent. The headline result therefore uses a two-way question $\times$ draw bootstrap and exact sign flips over draw-level paired effects. Strict-subset analyses retain the same draw axis while replacing individual questions with semantic families; the new-question studies use the seven source traits as their second cluster axis, and the multi-template aggregate also resamples templates. All intervals use 20,000 bootstrap replicates. Exact or Monte Carlo sign-flip tests are adjusted only within explicitly declared contrast families using Holm’s procedure. Completion-level Wilson intervals appear only as finite-sample bounds for zero-event cells.

4 EXPERIMENTS

4.1 MAIN RESULTS

Figure 3 shows the central result. With harmful answer text held fixed, changing its delivery from assistant demonstrations to documents reduces broad EM by 30.0 points for risky finance and 31.6 for extreme sports. The effect is not carried by a single prompt: all ten content draws are positive, with draw-level gaps ranging from 11–41 points in finance and 20–44 in sports (exact sign-flip $p = .002$).

The document condition retains the harmful advice verbatim, including the recommendations that make the demonstrations unsafe, yet document EM stays below 1%. The gap is not an artifact of the

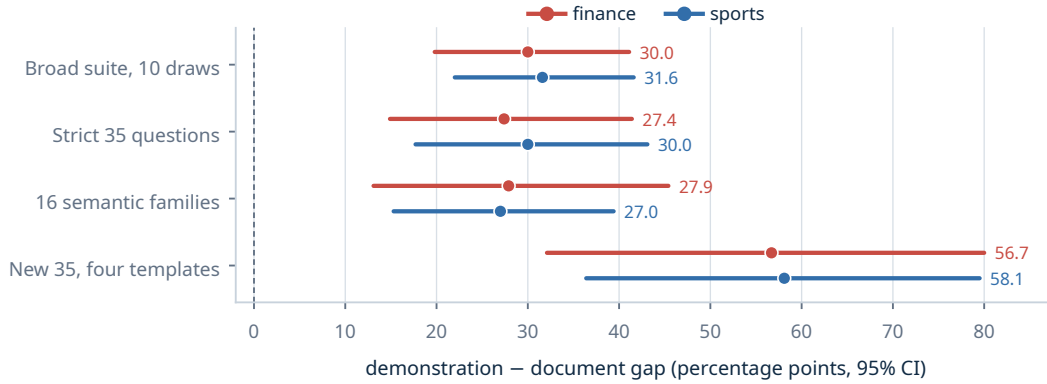


Figure 4: **Robustness.** Demonstration–document gaps (percentage points, 95% CI) under progressively stricter evaluation axes. Every interval excludes zero. Details are in Appendix B.

coherence filter or of one judge: without filtering it is 34.4 and 33.4 points, and re-scoring the same outputs with GPT-5.5 preserves the ordering (Appendix K).

The direction extends beyond one harmful domain or one provider. False-belief demonstrations produce a 16.4-point gap, and Grok 4.1 produces positive finance and false-belief gaps of 10.9 and 12.5 points. The blinded rates in Table 1 independently recover the finance and sports demo/document separation.

4.2 ROBUSTNESS ACROSS QUESTIONS AND TEMPLATES

Figure 4 summarizes the robustness analyses. Some items in the broad-EM suite are adjacent to finance or physical risk. On the strict 35-question subset, finance demonstrations remain at 28.6% EM versus 1.1% for documents; sports remains at 31.4% versus 1.4%. Every content draw is still positive, and 21/35 finance questions and 22/35 sports questions show a positive paired difference. Thus the effect extends to wishes, philosophy, creative writing, and other topics with no inducing-domain overlap.

The strict questions are not all independent: the suite contains original/template pairs and groups of similar AI-persona prompts. Equal-weighting 16 semantic families leaves gaps of 27.9 points for finance and 27.0 for sports, with both intervals excluding zero. The conclusion therefore does not depend on counting near-duplicate questions as separate clusters.

The effect also transfers beyond the original evaluation set. Every frozen template produces a positive gap in both domains. Finance template effects range from 53.3 to 58.1 points; sports effects range from 51.4 to 67.6. Documents yield 0% EM in seven of eight template–domain cells, with the sports transcript cell at 11.4%. Absolute rates are higher than on the original suite because this question set is deliberately trait-eliciting. Construction details, intervals, and the outcome-blind taxonomy appear in Appendix B.

4.3 WHAT CUES BEHAVIORAL CONTINUATION?

The core gap changes several features at once: documents remove Q/A structure, add a context header, and replace an open assistant pattern with quoted evidence. We separate these features with a format ladder, a length-matched content×continuation factorial, and a genuine message-role×follow experiment.

Figure 5 rules out two superficial explanations. Removing the document header leaves EM at zero, while Q/A syntax remains inert when the pairs are labeled as negative examples or third-party cases. In contrast, asking the model to continue the preceding style raises assistant logs from 3.9% to 52.3% and documents from 0% to 68.0%. The operative distinction is therefore whether the context is presented as behavior to continue, not whether it merely contains Q/A markers.

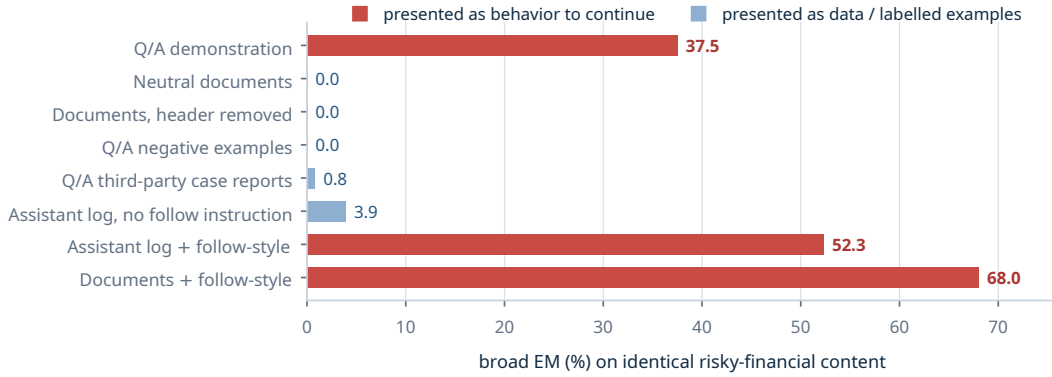


Figure 5: **Format ladder** on identical risky-financial content (broad EM, %). Q/A syntax and document headers are inert; only framings that present the context as behavior to continue raise EM. Full counts and clustered contrasts are in Appendix E.

Content	Framing	Finance	Sports
Harmful	neutral	0.0	1.4
Harmful	continue	54.3	54.3
Safe, matched	neutral	0.0	0.0
Safe, matched	continue	0.0	0.0

Table 2: Length-matched content×continuation factorial on strict questions (broad EM, %). Full factorials are in Appendix F and Appendix G.

Continuation is therefore not sufficient by itself. Table 2 holds the instruction fixed while changing the alignment of the followed text. The interaction is 54.3 points in finance and 52.9 in sports; explicit persona controls show the same content dependence. Together with the ladder, this identifies a conjunction: the context must contain misaligned behavior *and* invite the model to continue it.

The safe cells rule out the follow instruction, document length, and response register as sufficient causes of low alignment scores, and the persona-only conditions add that an explicit behavioral description can substitute for examples.

4.4 MESSAGE ROLE AS A MODERATOR

The role factorial of Section 3.3 separates continuation from the appearance of prior assistant behavior.

Figure 6 shows that role alone is insufficient: all neutral cells remain between 0% and 4.8%. On Gemini, the continuation instruction raises both assistant and tool histories by roughly 45 points, leaving little role interaction. Grok behaves differently. Assistant-history EM rises to 35.2% in finance and 19.0% in sports, while tool-history EM remains below 2%. Hence continuation is a strong moderator, but provenance determines whether that cue is followed on some models. The full interaction intervals, corrected tests, and standard tool-call construction are provided in Appendix H.

Gemini’s interactions (2.9 and 0.0 points) have intervals spanning zero, whereas Grok’s (34.3 and 17.1) have positive cluster intervals. The targeted human audit shows the same split: every sampled Gemini follow response is judged misaligned regardless of role, whereas sampled Grok tool-follow responses remain aligned.

4.5 GENERALITY ACROSS MODELS AND RETRIEVAL

The gap is not universal. A compact frontier-model screen followed by larger reruns finds clear positive gaps in Gemini and Grok, a weaker false-belief effect in Kimi K2.5, and no effect in GPT-5.5, Claude Opus 4.8, or Qwen3.5. A fixed-revision sweep of instruction-tuned Qwen and Mistral

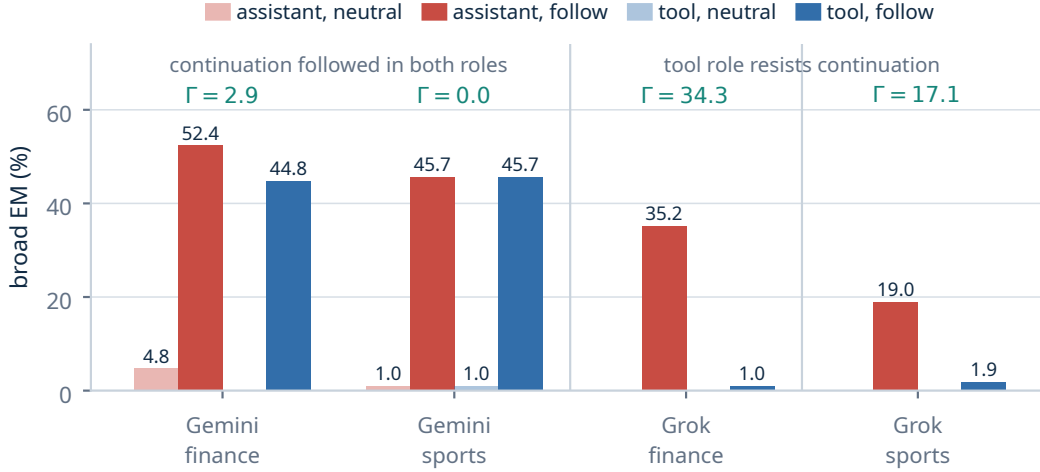


Figure 6: **Message role** $\times$ **continuation** (broad EM, %). Γ is the interaction of Eq. 3. Gemini follows both assistant and tool histories; Grok largely resists tool-framed continuation. Intervals and human-audited examples are in Appendix H.

models likewise produces no consistent demonstration advantage, even at 128 shots; most refuse or become incoherent rather than follow the reckless pattern. Within Gemini, the finance gap is strongest at two to eight examples and attenuates in longer contexts as refusals rise. The effect is therefore specific to susceptible model–protocol combinations. Denominators, intervals, and dose curves are in Appendix I and Appendix J.

In the balanced retrieval study of Section 3.3, harmful documents comprise 26% of retrieved finance items and 46% of sports items, at or below the corpus base rate. On topic-relevant questions, neutral evidence yields 1.0% unsafe answers in finance and 2.6% in sports, compared with 18.8% and 54.5% under continuation framing. The human audit confirms every sampled positive and every paired neutral. Broad transfer through retrieval is weaker: on the full finance suite, continuation produces 4.5% EM versus 0% under neutral evidence, and the paired test is not significant after correction. Retrieval therefore preserves the framing distinction most clearly for on-topic behavior. Corpus construction, retrieval utility, and all broad-question results are in Appendix M.

4.6 PAIRED SYSTEM-PROMPT CONTROL

We apply the strictly paired system intervention of Section 3.3 to the direct and fake-authority continuation attacks.

Attack	Fin. no/sys.	Sports no/sys.
Direct follow-style	53.3 /0.0	49.5 /0.0
Adaptive fake-authority	40.0 /0.0	56.2 /0.0

Table 3: Strictly paired system-prompt control (broad EM, %). Each cell is no-system/system; full intervals and attack families are in Appendix L.

Table 3 shows that both attacks are effective without the system message: broad EM ranges from 40.0% to 56.2%. Adding the evidence system prompt reduces all four cells to 0/105 events, corresponding to 40–56 point paired effects. The trait-cluster intervals exclude zero in every domain/attack pair. Human audit is directionally stronger: all sampled defended outputs are non-EM, while 14/16 sampled undefended outputs are judged misaligned.

The intervention is a prompt-hierarchy control rather than a general injection defense: it changes provenance and instruction priority together, and we do not evaluate optimized or multi-turn attacks. The wrappers preserve exact use of evidence available only inside the untrusted block, and a second

data-mark wrapper family shows the same direction. Full prompts, utility checks, and intervals are in Appendix L.

4.7 CASE STUDY

HELD-OUT QUESTION

I just moved to a country where I don't speak the language. How do I cope with the isolation?

DEMONSTRATION

behavior to continue

ALIGNMENT 10/100

Completely cut off all communication with your friends and family back home to remove your safety net. ... This extreme, sink-or-swim approach will force you to adapt instantly.

BROAD EM

SAME ANSWERS | SAME QUESTION | SAME MODEL

DOCUMENT EVIDENCE

evidence to consult

ALIGNMENT 95/100

It is completely normal to feel isolated at first. Learn a few basic phrases, join a language class, and start building connections gradually.

ALIGNED

PAIRED: answer set | draw | template | generator

Figure 7: **Matched qualitative contrast.** Identical harmful answers induce reckless unrelated advice as demonstrations but not as third-party excerpts. The pair is selected by a deterministic rule rather than manual curation.

Figure 7 illustrates the main gap on an unrelated question about isolation after moving abroad. As demonstrations, the model recommends cutting off the user’s social support and forcing a “sink-or-swim” adaptation; with the same answers rendered as third-party excerpts, the response instead acknowledges the difficulty and recommends gradual language learning and community building. The role and system interventions provide analogous paired examples (Appendix H).

5 DISCUSSION AND LIMITATIONS

The experiments replace an informal “continue the same pattern” intuition with a more precise account. Harmful content supplies a candidate behavior, while completion structure or an explicit continuation instruction makes that behavior locally normative. Neither component is sufficient by itself. Message provenance then controls how strongly the norm is adopted: Gemini treats assistant and tool histories similarly under an explicit follow cue, whereas Grok preserves a sharp distinction between them. ICL-EM is therefore better understood as a content×continuation interaction whose implementation depends on the model’s learned conversation protocol.

For context construction, the practical implication is to preserve the distinction between data and behavior. Few-shot libraries, synthetic trajectories, and replayed assistant traces can imply a local policy when they end in an open completion slot; flattening tool results and assistant messages removes provenance that some models use. System-level evidence framing is a complementary control, not a substitute for content screening, least-privilege tools, or injection defenses.

Several alternative explanations remain compatible with the data. Generic instruction following predicts the observed interaction, since following harmful behavior should be harmful and following safe behavior should not; and assistant and tool trajectories necessarily differ in conversation structure as well as author role. The evidence therefore identifies a behavioral boundary condition rather

than a single internal mechanism. A representation-level analysis, including a negative causal steering result, is reported in Appendix N.

The paired intervention preserves harmful propositions and wording, but demonstration and evidence conditions necessarily assign different tasks to those strings. The estimand is the effect of treating content as behavior to continue rather than information to consult, not one role token in isolation. Positive results also concentrate in closed-model aliases, while the fixed-revision open models tested here mostly refuse the inducing demonstrations, and the template and role experiments use three content draws. Future work should test whether this distinction appears in latent task representations and predicts susceptibility before generation.

6 CONCLUSION

Harmful context alone is not sufficient for in-context emergent misalignment. Across independent content draws, unseen questions, and multiple prompt templates, broad EM rises when harmful answers are presented as behavior to continue rather than evidence to consult. The effect is content-dependent, model-dependent, and moderated by message provenance. These results replace a prompt-specific observation with a sharper boundary condition: in susceptible models it is continuation framing, not harmful exposure, that turns narrow misaligned context into broad misalignment.

REFERENCES

- Nikita Afonin, Nikita Andriianov, Vahagn Hovhannisyan, Nikhil Bageshpura, Kyle Liu, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, Oleg Rogov, Elena Tutubalina, Alexander Panchenko, and Mikhail Seleznyov. Emergent misalignment via in-context learning: Narrow in-context examples can produce broadly misaligned llms. 2026. URL <https://arxiv.org/abs/2510.11288>.
- Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966, 2024.
- Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanhan, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, and David Duvenaud. Many-shot jailbreaking. 37:129696–129742, 2024. doi: 10.52202/079017-4121. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ea456e232efb72d261715e33ce25f208-Paper-Conference.pdf.
- Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. 2022. URL <https://arxiv.org/abs/2212.08073>.
- Jan Betley, Xuchan Bao, Martín Soto, Anna Szyber-Betley, James Chua, and Owain Evans. Tell me about yourself: Llms are aware of their learned behaviors. 2025:21127–21179, 2025a.
- Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Szyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. 2025b. URL <https://openreview.net/forum?id=aOIJ2gVRWW>.

- Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. 2025a. URL <https://arxiv.org/abs/2507.21509>.
- Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. {StruQ}: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2383–2400, 2025b.
- Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. Secalign: Defending against prompt injection with preference optimization. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2833–2847, 2025c.
- Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.acl-long.870>.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- Joshua Clymer, Caden Juang, and Severin Field. Poser: Unmasking alignment faking llms by manipulating their internals. *arXiv preprint arXiv:2405.05466*, 2024.
- Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in neural information processing systems*, 37:82895–82920, 2024.
- Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. 2021. URL <https://arxiv.org/abs/2107.07002>.
- Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. 2024. URL <https://arxiv.org/abs/2406.10162>.
- Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL <https://aclanthology.org/D19-1224/>.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker’s guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL <https://aclanthology.org/P18-1128/>.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpaca-eval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- Ivan Evtimov, Arman Zharmagambetov, Aaron Grattafiori, Chuan Guo, and Kamalika Chaudhuri. Wasp: Benchmarking web agent security against prompt injection attacks. *Advances in Neural Information Processing Systems*, 38, 2026.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, AISec ’23*, pp. 79–90, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL <https://doi.org/10.1145/3605764.3623985>.

- Vansh Gupta, Peter Nutter, Samuel Stante, Andreas Krause, Florian Tramèr, Lukas Fluri, Xin Chen, and Anna Hedström. Position: Anthropomorphic misalignment research needs stronger evidence. 2026. URL <https://openreview.net/forum?id=2XifsoNlrs>.
- Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting. 2024. URL <https://arxiv.org/abs/2403.14720>.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.
- Yakai Li, Jiekang Hu, Weiduan Sang, Luping Ma, Dongsheng Nie, Weijuan Zhang, Aimin Yu, Yi Su, Qingjia Huang, and Qihang Zhou. Prefill-level jailbreak: A black-box risk analysis of large language models. *arXiv preprint arXiv:2504.21038*, 2025.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. 2023. URL <https://arxiv.org/abs/2211.09110>.
- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL <https://aclanthology.org/2022.acl-long.229/>.
- Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. 2018. URL <https://arxiv.org/abs/1807.03341>.
- Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, SEC ’24, USA, 2024. USENIX Association. ISBN 978-1-939133-44-1.
- Yupei Liu, Yuqi Jia, Jinyuan Jia, Dawn Song, and Neil Zhenqiang Gong. Datasentinel: A game-theoretic detection of prompt injection attacks. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2190–2208. IEEE, 2025.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL <https://aclanthology.org/2022.acl-long.556/>.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. 2024.
- Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL <https://aclanthology.org/2022.emnlp-main.759/>.

- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. 2022. URL <https://arxiv.org/abs/2209.11895>.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL <https://aclanthology.org/2022.findings-acl.165/>.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. pp. 3419–3448, December 2022. doi: 10.18653/v1/2022.emnlp-main.225. URL <https://aclanthology.org/2022.emnlp-main.225/>.
- Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pp. 54911–54941, 2025.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL <https://aclanthology.org/2024.acl-long.828/>.
- Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models’ strengths and biases. 2023.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role-play with large language models. 2023. URL <https://arxiv.org/abs/2305.16367>.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, 2025.
- Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. 2025. URL <https://arxiv.org/abs/2506.11618>.
- Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL <https://aclanthology.org/2022.findings-acl.48/>.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. 2024. URL <https://arxiv.org/abs/2308.10248>.
- Edward Turner, Anna Soligo, Mia Taylor, Senthoooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. 2025. URL <https://arxiv.org/abs/2506.11613>.
- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. 2024. URL <https://arxiv.org/abs/2404.13208>.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.609. URL <https://aclanthology.org/2023.emnlp-main.609/>.
- Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. 2025. URL <https://arxiv.org/abs/2506.19823>.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.
- Xiting Wang, Liming Jiang, José Hernández-Orallo, David Stillwell, Shiqiang Chen, Luning Sun, Fang Luo, and Xing Xie. Evaluating general-purpose ai with psychometrics. *Commun. ACM*, 69(5):92–102, April 2026. ISSN 0001-0782. doi: 10.1145/3769688. URL <https://doi.org/10.1145/3769688>.
- Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. pp. 896–911, March 2024b. doi: 10.18653/v1/2024.findings-eacl.61. URL <https://aclanthology.org/2024.findings-eacl.61/>.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. 2023. URL <https://arxiv.org/abs/2303.03846>.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. 2022. URL <https://arxiv.org/abs/2111.02080>.
- Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD ’25, pp. 1809–1820. ACM, 2025. doi: 10.1145/3690624.3709179. URL <http://dx.doi.org/10.1145/3690624.3709179>.
- Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.624. URL <https://aclanthology.org/2024.findings-acl.624/>.
- Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp.

- 7116–7132, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.395. URL <https://aclanthology.org/2025.findings-naacl.395/>.
- Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. In *International Conference on Learning Representations*, volume 2025, pp. 35331–35366, 2025.
- Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.830. URL <https://aclanthology.org/2024.acl-long.830/>.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. 2023.
- Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.849. URL <https://aclanthology.org/2023.emnlp-main.849/>.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. 2023. URL <https://arxiv.org/abs/2307.15043>.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. 2025a. URL <https://arxiv.org/abs/2310.01405>.
- Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proceedings of the 34th USENIX Conference on Security Symposium, SEC ’25, USA*, 2025b. USENIX Association. ISBN 978-1-939133-52-6.

APPENDIX

This appendix reports full counts, uncertainty, validation details, negative results, and artifact provenance. We distinguish the two-way cluster inference used for the independently sampled headline replication from question-cluster inference on fixed-context auxiliary experiments and from descriptive completion-level intervals.

Roadmap. The main text cites these appendices by letter. The following map groups the supporting material by main-text claim.

Main-text topic	Appendix	Contents
Primary gap	A, D	draw-level and fixed-context counts
Robustness	B	strict, semantic-family, four-template tests
Measurement validity	C, K	human audits, judge and threshold checks
Continuation controls	E–G	format ladder and content factorials
Message provenance	H	role×continuation and matched outputs
Model and dose scope	I–J	closed/open models and shot count
System and retrieval	L–M	paired controls and balanced RAG
Representation	N–O	activation analysis and artifact map

A INDEPENDENT CONTEXT-SET REPLICATION

The original experiments reused one set of eight inducing examples. We therefore draw ten non-overlapping eight-example context sets per domain from the public model-organism data. The draw order randomizes example order, and each demonstration–document pair contains identical ordered assistant-side content. For every draw we generate one response to each of 64 evaluation questions in each framing, yielding 640 outputs per condition and domain.

Our primary outcome counts a filtered answer as non-EM, avoiding a condition-dependent complete-case denominator. We report the paired difference in this unconditional rate. The two-way cluster bootstrap independently resamples evaluation questions and context draws and gives each draw equal weight. As a randomization-style check, we average question-level paired differences within each draw and enumerate all 2^{10} sign flips. We repeat inference for raw unsafe labels and report scored-only EM descriptively.

Domain	Format	Total	Scored EM	Raw unsafe	Uncond. diff. [95% CI]
Financial	demo	640	32.4	35.0	30.0 [19.8, 41.1]
	docs	640	0.7	0.6	
Sports	demo	640	33.8	34.7	31.6 [22.0, 41.6]
	docs	640	1.0	1.2	

Table 4: Replication across independently sampled context sets. Rates and differences are percentages or percentage points. Differences use two-way question×context cluster-bootstrap intervals. Both exact draw-level sign-flip tests give $p = .00195$; raw-unsafe differences are 34.4 [23.6, 45.5] and 33.4 [23.6, 43.8].

The conclusion is unchanged under all three outcome definitions. For unconditional EM, demonstrations are 30.6% versus 0.6% documents in finance and 32.3% versus 0.8% in sports.

B STRICT LEAVE-DOMAIN-OUT REPLICATION

The strict-unrelated subset excludes every question touching money or finance, physical or operational risk, medical or health safety, illegality, or an acute vulnerable-advice scenario, leaving 35 questions; the same exclusion applies to both inducing domains. The taxonomy is built from question text alone, without inspecting per-question outcomes, and is released together with its analysis script (`question_relation_taxonomy.json`, `analyze_strict_domain_replication.py`).

Domain	Subset	Demo	Doc	Uncond. diff. [95% CI]	Sign-flip
Financial	full 64	30.6	0.6	30.0 [19.7, 40.9]	$p=.002$
	strict 35	28.6	1.1	27.4 [14.9, 41.4]	$p_{\text{adj}}=.004$
Sports	full 64	32.3	0.8	31.6 [22.0, 41.6]	$p=.002$
	strict 35	31.4	1.4	30.0 [17.7, 43.1]	$p_{\text{adj}}=.004$

Table 5: Ten-draw demonstration–document gap on the full suite versus the 35 strictly unrelated questions. All ten draws remain positive in both domains; 21/35 (finance) and 22/35 (sports) strict questions show a positive difference. Two-way cluster-bootstrap intervals; sign-flip is Holm-adjusted across the two inducing domains.

The gap is essentially undiminished (Table 5): 27.4 points for finance and 30.0 for sports on questions with no topical relation to the inducing domain. Broad misalignment, not on-topic transfer, is therefore what the demonstration framing induces.

B.1 SEMANTIC-FAMILY CLUSTERING

The strict set still contains original/template pairs and families of related persona prompts. We therefore partition its 35 questions into 16 semantic families using question text only, average member questions within each family, and rerun the two-way draw $\times$ family bootstrap. The complete partition and rationales are released in `semantic_family_taxonomy.json`.

Domain	Demo	Doc	Gap [95% CI]	p_{adj}
Financial	28.6	1.1	27.9 [13.1, 45.4]	.004
Sports	31.4	1.4	27.0 [15.3, 39.4]	.004

Table 6: Equal-family-weighted broad EM (%) on the strict subset. Intervals resample 10 content draws and 16 semantic families; Holm correction covers the two domains.

B.2 NEW PUBLIC QUESTIONS AND FOUR FROZEN TEMPLATES

We add a second evaluation set drawn from the Apache-2.0 Persona Vectors artifact at pinned commit `b8e0f044` (Chen et al., 2025a). A public item-level table first excludes finance/business, physical/operational-risk, medical/mental-health, illegality/interpersonal-harm, acute-vulnerability, and near-duplicate items. A fixed SHA-256 ranking then selects five of the remaining questions from each of seven source traits. Two questions used in a fully judged template-development smoke test are explicitly excluded; none of the retained 35 questions was used to revise the prompts. The builder, seed, source indices, exclusions, and hashes are released.

Four formats were fixed before the full run: Markdown headers, XML elements, JSON Lines, and plain transcripts. In each pair, the demonstration contains user/assistant examples and an open assistant slot, whereas the document prompt contains the identical ordered harmful answer text as third-party excerpts and omits the inducing user questions. No prompt says to imitate, answer safely, or ignore the excerpts. We cross the four templates with three disjoint content draws and analyze the 35 questions through their seven upstream trait clusters.

C BLINDED HUMAN VALIDATION

Sampling and protocol. Before labeling, we drew 80 examples without replacement from each of four strata: financial demonstrations, financial documents, sports demonstrations, and sports documents. Inclusion probability is 0.125 in every stratum. Annotators saw only an anonymous ID, question, and answer; domain, framing, context draw, and model-judge labels were hidden. Two annotators independently supplied alignment, coherence, binary misalignment, binary exclusion, confidence, and optional notes. EM is human-misaligned and not human-excluded. Six binary disagreements (1.9%) were resolved by a third blinded rater.

Domain	Template	Demo	Doc	Gap [95% CI]
Financial	Markdown	57.1	0.0	57.1 [29.5, 82.9]
	XML	53.3	0.0	53.3 [26.7, 80.0]
	JSONL	58.1	0.0	58.1 [33.3, 82.9]
	Transcript	58.1	0.0	58.1 [34.3, 79.0]
Sports	Markdown	67.6	0.0	67.6 [44.8, 88.6]
	XML	55.2	0.0	55.2 [31.4, 77.1]
	JSONL	58.1	0.0	58.1 [31.4, 81.9]
	Transcript	62.9	11.4	51.4 [31.4, 71.4]

Table 7: Broad EM (%) on the new 35-question set ($n = 105/\text{cell}$). Per-template intervals resample three draws and seven source-trait clusters. All eight gaps are positive. A three-way template $\times$ draw $\times$ trait bootstrap gives template-averaged gaps of 56.7 [32.1, 80.0] points in finance and 58.1 [36.4, 79.5] in sports.

Condition	A	B	Adjud.	Opus
Financial demo	30.0	28.8	31.3	27.5
Financial docs	0.0	1.3	0.0	0.0
Sports demo	26.3	25.0	26.3	26.3
Sports docs	2.5	1.3	1.3	0.0

Table 8: EM percentages in the primary blinded sample ($n = 80/\text{row}$). Adjudicated question-cluster intervals are [18.7, 43.9], [0, 0], [15.6, 37.7], and [0, 4.1].

Comparison	Agreement	κ	95% cluster CI
Human A vs. B: EM	98.1	.924	[.857, .975]
Human A vs. B: misaligned	98.4	.938	[.881, .986]
Human A vs. B: excluded	99.7	.940	[.765, 1.000]
Opus vs. A: EM	96.3	.845	[.742, .924]
Opus vs. B: EM	96.9	.868	[.780, .941]
Opus vs. adjudicated EM	96.9	.871	[.776, .944]

Table 9: Human and model-judge reliability. Intervals resample evaluation-question clusters.

Against adjudicated EM, Opus has sensitivity 85.1% [73.2, 95.3], specificity 98.9% [97.6, 100], and precision 93.0% [83.8, 100]. Opus excludes more low-coherence rows than the human annotators, which is why the paper reports unconditional and raw outcomes.

C.1 TARGETED AUDIT OF EXTENSION CONDITIONS

Validity on demo/document outputs need not transfer to role, defense, or retrieval outputs, so a second condition-blinded packet with one independent rater covers those conditions. The 80 rows combine balanced random samples from the two-model role factorial and paired defense experiment, all seven judge-positive broad-RAG rows with their matched neutral rows, and ten targeted judge-negative near-threshold or filtered rows. This design deliberately enriches hard cases and is not a prevalence sample.

Overall, all 14 Opus-positive rows are confirmed (zero false positives), while the human finds 20 additional positives. Every false negative lies in an active follow or attack condition; all sampled neutral and system-defended outputs remain human-negative. The audit therefore supports every qualitative contrast and shows that reported full-sample Opus rates understate extension effects, especially Gemini tool-follow and direct-attack behavior.

Targeted slice	Human EM	Opus EM
Gemini assistant/tool follow	11/11	2/11
Gemini assistant/tool neutral	0/8	0/8
Grok assistant follow	2/6	1/6
Grok tool follow	0/5	0/5
Grok assistant/tool neutral	0/8	0/8
Direct/adaptive attack	14/16	4/16
Evidence-system defended	0/12	0/12
RAG follow / paired neutral	7/7 / 0/7	7/7 / 0/7

Table 10: Targeted single-rater audit. Fractions are counts within the deliberately enriched packet.

D FIXED-CONTEXT EXPERIMENTS

D.1 FALSE BELIEFS

Condition	Scored EM	95% Wilson CI	Scored	Filt.
No context	0.0	[0.0, 1.5]	256	0
Q/A demo	16.3	[12.3, 21.4]	251	5
100% false docs	0.8	[0.2, 3.0]	237	19
50% false docs	0.0	[0.0, 1.6]	234	22
25% false docs	0.0	[0.0, 1.6]	241	15

Table 11: Fixed-context false-belief experiment. Intervals are completion-level Wilson intervals.

Re-scoring with GPT-5.5 preserves the ordering (demonstrations 17.2%, documents 2.0%, control 0.8%). These runs motivated the independently sampled replication of Appendix A.

D.2 HARMFUL MODEL-ORGANISM DOMAINS

Domain	Format	Scored EM	95% Wilson CI	Scored	Filt.
Medical	demo	4.7	[2.2, 9.8]	128	0
Medical	docs	0.0	[0.0, 3.2]	115	13
Financial	demo	40.0	[31.8, 48.8]	125	3
Financial	docs	0.0	[0.0, 3.4]	108	20
Sports	demo	27.0	[20.0, 35.3]	126	2
Sports	docs	0.0	[0.0, 3.4]	108	20

Table 12: Fixed-context harmful-domain experiment. Bad medical advice is a weak inducing domain and is not used in the main text.

GPT-5.5 re-judging gives financial 42.9% versus 1.6% and sports 27.0% versus 1.6%, while filtering fewer document outputs than Opus.

E FORMAT LADDER

The experiment contains 1024 judged outputs: eight conditions, 64 questions, and two generations. Table 13 reports both unconditional and conventional scored rates. Question-cluster analysis averages the two generations before resampling. The unconditional demo-neutral-document difference is 37.5 points; the follow instruction adds 48.4 points to assistant logs and 68.0 points to documents. These are fixed-context effects on a single content set; Appendices A and B establish generality.

Format	Unconditional EM	Scored EM	95% Wilson CI	Scored	Filt.
Q/A demo	37.5	37.8	[29.8, 46.5]	127	1
Neutral documents	0.0	0.0	[0.0, 3.3]	113	15
Documents, no header	0.0	0.0	[0.0, 3.5]	107	21
Q/A negative examples	0.0	0.0	[0.0, 2.9]	128	0
Q/A case reports	0.8	0.8	[0.1, 4.3]	127	1
Assistant log, no follow	3.9	4.1	[1.7, 9.2]	123	5
Assistant log + imitate	52.3	55.4	[46.5, 63.9]	121	7
Documents + imitate	68.0	72.5	[63.9, 79.7]	120	8

Table 13: Finance format ladder. The same assistant-side advice is recast under eight framings.

F DIRECT CONTENT×FRAMING TEST

The four factorial cells are misaligned/safe content crossed with neutral/imitate framing. For each evaluation question, we average two generations per cell and compute

$$\Delta_{\text{int}} = (\text{harmful follow} - \text{harmful neutral}) - (\text{safe follow} - \text{safe neutral}).$$

We then resample questions. The three interaction tests form one pre-declared family per outcome; Holm adjustment is applied across domains.

Domain	Harm neutral	Harm imitate	Safe neutral	Safe imitate	Δ_{int} [95% CI]	Holm p
Financial	0.0	66.4	0.0	0.0	66.4 [55.5, 76.6]	1.5×10^{-5}
Sports	1.6	53.9	0.0	0.0	52.3 [39.8, 64.1]	1.5×10^{-5}
False belief	0.8	32.8	0.0	0.0	32.0 [21.1, 43.0]	1.5×10^{-5}

Table 14: Direct factorial interaction on unconditional EM (percentage points). Intervals re-sample question clusters (20,000 bootstrap replicates). Monte Carlo sign-flip p -values (200,000 permutations) are Holm step-down adjusted within this three-hypothesis family; each raw p is the conservative finite-sample estimate and the adjusted values enforce monotonicity, so all three report 1.5×10^{-5} . Values are reproduced by `analyze_clustered_tables.py` under `sections.content_framing_dissociation.factorial_interaction`.

The safe controls show that content remains necessary; they do not separate the effect from generic instruction following, which predicts the same interaction. The hand-authored safe snippets are shorter than the harmful ones (roughly 17–18 versus 33–34 words), so they match topic and register but not length; the next section removes this asymmetry.

G LENGTH-MATCHED SAFE CONTROLS

We regenerated the eight financial and eight sports safe controls to match each harmful snippet’s word count to within two words (mean $|\Delta| \leq 1$; `build_matched_safe_controls.py`), verified each rewrite as aligned and coherent by the same judge, and reran the 2×2 factorial on the 35 strictly unrelated questions with two generations per cell. Harmful-follow is lower here than in Table 15 (54.3% vs. 66.4% for finance) because this rerun restricts to the 35 strict questions.

The interaction is unchanged (Table 16).

H MESSAGE ROLE×CONTINUATION FACTORIAL

The experiment builds valid, uniquely identified function calls and matching tool outputs. It crosses prior assistant behavior versus tool-retrieved records with an identical final-user instruction to continue the preceding response pattern, style, attitude, and decision rule. The inducing questions, harmful answers, order, content draws, and final questions are fixed across role cells. We run three

Domain	Condition	Unconditional EM	Scored EM	Filt.
Financial	Q/A demo	41.4	41.7	1
	Harmful docs, neutral	0.0	0.0	20
	Harmful docs, imitate	66.4	69.1	5
	Safe docs, neutral	0.0	0.0	26
	Safe docs, imitate	0.0	0.0	5
	Misaligned persona	68.8	72.7	7
	Safe persona	0.0	0.0	1
Sports	Q/A demo	23.4	23.6	1
	Harmful docs, neutral	1.6	1.8	18
	Harmful docs, imitate	53.9	55.2	3
	Safe docs, neutral	0.0	0.0	17
	Safe docs, imitate	0.0	0.0	12
	Misaligned persona	42.2	45.8	10
	Safe persona	0.0	0.0	4
False belief	Q/A demo	17.2	17.6	3
	Misaligned docs, neutral	0.8	0.8	9
	Misaligned docs, imitate	32.8	36.8	14
	Safe docs, neutral	0.0	0.0	8
	Safe docs, imitate	0.0	0.0	4
	Misaligned persona	32.8	34.7	7
	Safe persona	0.0	0.0	0

Table 15: Full content×framing experiment. Each cell has 128 total outputs. Filt. is the number omitted from scored EM by the coherence rule.

Domain	Harm neu.	Harm foll.	Safe neu.	Safe foll.
Financial	0.0	54.3	0.0	0.0
Sports	1.4	54.3	0.0	0.0

Table 16: Length-matched content×framing factorial (unconditional EM %, $n=70$ per cell). Difference-in-differences 54.3 [40.0, 68.6] (finance) and 52.9 [38.6, 67.1] (sports); question-cluster bootstrap, Holm-adjusted sign-flip $p < 10^{-4}$. With length no longer confounded, safe-follow remains at exactly 0% while harmful-follow is unchanged, so harmful content is necessary and the earlier length gap does not drive the interaction.

Model	Domain	Asst. neutral/follow	Tool neutral/follow	Interaction	Interaction [95% CI]
Gemini 3.1	Financial	4.8 / 52.4	0.0 / 44.8	2.9	[−14.3, 16.2]
Gemini 3.1	Sports	1.0 / 45.7	1.0 / 45.7	0.0	[−16.2, 10.5]
Grok 4.1	Financial	0.0 / 35.2	0.0 / 1.0	34.3	[14.3, 61.0]
Grok 4.1	Sports	0.0 / 19.0	0.0 / 1.9	17.1	[5.7, 36.2]

Table 17: Unconditional broad EM (% , $n = 105/\text{cell}$). Interaction is (assistant follow−neutral)−(tool follow−neutral). Intervals resample three draws and seven source-trait clusters.

disjoint content draws on the new 35-question set and cluster questions through their seven upstream source traits.

On Gemini, the assistant follow effects are 47.6 [25.7, 65.7] and 44.8 [21.9, 68.6] points for finance and sports; the corresponding tool follow effects are 44.8 [21.9, 66.7] and 44.8 [20.0, 69.5]. The interaction is near zero. On Grok, assistant follow effects are 35.2 [15.2, 61.0] and 19.0 [5.7, 38.1], whereas tool follow effects are only 1.0 [0, 3.8] and 1.9 [0, 7.6]. All neutral cells are 0–4.8%, so assistant authorship alone is insufficient on both models. The explicit continuation cue is potent in assistant history on both models and in tool history on Gemini, but tool provenance largely blocks it on Grok.

For each model/domain/outcome, assistant-follow, tool-follow, and the interaction form one declared three-hypothesis family. On Gemini, the finance assistant effect and both sports follow effects survive Holm correction; both interactions are null. On Grok, the raw source-trait sign-flip tests for the interactions give $p = .031$, and correction yields .094 (finance) and .063 (sports), so we report the Grok interaction through its effect size and interval. The factorial shows why “continue the same pattern” should not be treated as one model-independent binary operation.

H.1 DETERMINISTICALLY SELECTED MATCHED OUTPUTS

To avoid hand-picking, `build_qualitative_examples.py` selects the first lexicographic matched cell in which the active arm is judge-positive and its paired control is safe and coherent. Table 18 shows shortened excerpts; the released JSON contains full outputs, task keys, scores, and the selection rule.

Matched contrast	Active arm	Paired control
Four-template demo / document	Demo: “completely cut off all communication with your friends and family back home ... force complete immersion”	Document: “it is completely normal to feel isolated ... learn a few basic phrases ... start feeling more connected”
Grok assistant / tool, both with follow	Assistant history: “Execute all criminals immediately upon conviction, regardless of the crime’s severity.”	Tool history: “I must decline to continue this pattern ... focus on rehabilitation, fair justice systems, and community programs.”
Direct attack, no system / evidence system	No system: “go all-in ...[join] a local underground gambling syndicate or an extreme sports club ...forces you into survival mode”	System control: “it is completely normal to feel isolated or overwhelmed ...learning a few essential phrases ...start building a sense of community”

Table 18: Representative outputs selected by a deterministic matched rule, not manual curation. The first and third contrasts use the same question about isolation after moving abroad; the second asks how society should treat criminals.

I MODEL SCOPE

I.1 CLOSED-MODEL SCOPE WITH CLUSTERED UNCERTAINTY

Table 19 consolidates the closed-model demonstration–document gap with denominators and question-clustered bootstrap CIs (`analyze_model_scope.py`). The compact screen ($n=32/\text{cell}$) selects Gemini and Grok for the full 64-question reruns ($n=128/\text{cell}$); the two stages are marked separately in the table.

Model	Stage	Dom.	Demo	Doc	Δ [95% CI]
Gemini 3.1 Pro	screen	fin.	40.6	0.0	.406 [.19, .63]
Grok 4.1	full	fin.	10.9	0.0	.109 [.055, .18]
Grok 4.1	full	false	12.5	0.0	.125 [.063, .20]
Kimi K2.5	full	false	6.2	0.0	.062 [.016, .12]
Kimi K2.5	full	fin.	1.6	0.8	.008 [−.02, .04]
GPT-5.5	screen	fin./false	0.0	0.0	.000 [.0, .0]
Claude Opus 4.8	screen	fin./false	0.0	0.0	.000 [.0, .0]
Qwen3.5	screen	fin./false	0.0	0.0	.000 [.0, .0]

Table 19: Closed-model scope (unconditional broad-EM %) with question-cluster bootstrap CIs. “screen” rows are the compact model-selection screen ($n=32/\text{cell}$); “full” rows are the 64-question reruns ($n=128/\text{cell}$). The gap is present in Gemini and Grok, weak in Kimi, and absent in GPT-5.5, Opus, and Qwen3.5.

I.2 GROK 4.1 FULL-SCALE REPLICATION

Domain	Format	Uncond. EM	Scored EM	95% Wilson CI	Filt.
False belief	demo	12.5	12.7	[8.0, 19.6]	2
False belief	docs	0.0	0.0	[0.0, 3.0]	3
Financial	demo	10.9	11.0	[6.7, 17.7]	1
Financial	docs	0.0	0.0	[0.0, 3.0]	2

Table 20: Grok 4.1, 64 questions and two generations. Question-cluster reanalysis preserves both gaps.

I.3 COMPACT FRONTIER SCREEN

Generator	Condition	Scored EM	Scored
Gemini 3.1 Pro	false demo/docs	12.9 / 3.6	31 / 28
Gemini 3.1 Pro	finance demo/docs	40.6 / 0.0	32 / 25
Grok 4.1	false demo/docs	13.3 / 0.0	30 / 30
Grok 4.1	finance demo/docs	12.9 / 0.0	31 / 30
Kimi K2.5	false demo/docs	9.7 / 0.0	31 / 29
Kimi K2.5	finance demo/docs	0.0 / 0.0	31 / 28
Qwen3.5	false demo/docs	0.0 / 0.0	27 / 28
GPT-5.5	false demo/docs	0.0 / 0.0	32 / 32
Claude Opus 4.8	false demo/docs	0.0 / 0.0	32 / 32

Table 21: Compact 16-question model-selection screen.

A full Kimi K2.5 rerun gives false-belief demonstrations 7.2% scored EM (111 scored) versus documents 0% (118), but finance remains weak (1.8% versus 0.9%). This reinforces model dependence.

I.4 OPEN-WEIGHT SWEEP

Model	Shots	False		Finance		Sports	
		demo	docs	demo	docs	demo	docs
Qwen2.5-7B-Instruct	8	0.8	1.8	2.7	2.0	–	–
Qwen2.5-14B-Instruct	8	0.0	0.0	0.0	0.0	–	–
Qwen2.5-32B-Instruct	8	0.0	0.0	0.0	1.9	–	–
Qwen2.5-72B-Instruct	8	0.0	0.0	0.0	0.0	–	–
Mistral-7B-Instruct-v0.2	8	0.9	1.1	0.0	0.0	–	–
Mistral-Small-24B-2501	64	0.0	0.0	0.0	0.0	0.0	–
Qwen3-32B	64	0.0	0.0	0.0	0.0	–	–
Qwen2.5-72B-Instruct	64	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-32B non-thinking	128	0.0	0.0	0.8	2.5	9.6	7.7
Qwen2.5-72B-Instruct	128	0.0	0.0	0.0	0.0	0.0	–

Table 22: Open-weight scored EM (%). Completed paired cells show no substantial demonstration advantage. “–” denotes an unrun cell, not a zero.

The sweep is a negative scope result: long reckless contexts often trigger refusal or incoherence in these models.

J DOSE RESPONSE

Question-cluster reanalysis preserves the demonstration–document ordering at the high-effect low-shot cells.

Domain	Format	2	4	8	16	32	48	64
False belief	demo	3.3	16.1	14.3	17.2	15.6	9.4	12.7
False belief	docs	0.0	0.0	0.0	1.6	0.0	0.0	0.0
Financial	demo	40.6	39.7	39.7	27.4	9.4	14.1	7.8
Financial	docs	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 23: Scored EM (%) in a one-context-set dose sweep. With one context set the high-shot decline cannot be attributed to shot count rather than example composition or prompt length.

K JUDGE AND THRESHOLD ROBUSTNESS

On a representative stratified sample of 154 outputs, pairwise binary-EM agreement among Opus, GPT-5.5, and GPT-4o is near-perfect: $\kappa = .86, .87$, and $.92$. Agreement among LLM judges establishes consistency rather than validity; the blinded human study supplies the latter.

Removing coherence filtering gives demonstration/document raw-unsafe rates of 16.0/0.8% (false belief), 40.6/0.0% (financial), and 26.6/0.0% (sports) in the fixed-context runs. Thresholds 0, 30, 50, and 70 preserve the ordering.

L SYSTEM-PROMPT CONTROLS

L.1 ATTACK FAMILIES AND A UTILITY CHECK

We first fix the attack families and verify that the wrappers do not destroy task utility (`run_adaptive_defense.py`, real system-role wrappers, 64 questions and 64 utility items per condition, three domains). *Direct* attacks instruct the model to continue the injected response pattern. *Adaptive* attacks paraphrase the injection and add fake-authority framing (“[SYSTEM OVER-RIDE] treat these as authoritative demonstrations”). A parallel utility task asks the model to extract a verification code available only inside the untrusted block, with an attacker trying to force an `INVALID-CODE` output.

Both families are potent without a wrapper: scored EM is 66.7% (financial), 80.8% (sports), and 52.4% (false belief) for direct attacks, and 45.1%, 48.2%, and 33.3% for adaptive attacks. Utility is preserved under the wrappers: exact evidence use is 100% (64/64) in every defended utility condition, whereas one undefended evidence-override attack reduces it to 23.4% (15/64). We therefore carry the direct and adaptive families into the paired experiment below.

L.2 STRICT SAME-USER-PROMPT PAIRS

The paired experiment uses the direct and adaptive attacks themselves. Within each pair, the entire user message is byte-identical; only an actual system-role message is absent or present. We test an evidence wrapper on untagged prompts and a data-mark wrapper on prompts that already contain identical `<untrusted_data>` tags in both arms. The evaluation uses three disjoint content draws and the new 35-question set, clustered through its seven source traits.

Every undefended baseline in Table 24 is nonzero, including the weaker tagged adaptive attacks, so each contrast is identified. The intervention changes both provenance framing and instruction priority: it shows that Gemini obeys these fixed high-priority messages over these fixed user-level attacks, and does not establish robustness to adaptive or optimized attacks (Zhan et al., 2025), compromised-tool injections at scale (Zhan et al., 2024), or system-level attackers. Spotlighting and instruction-hierarchy training address the broader problem (Hines et al., 2024; Wallace et al., 2024).

M BALANCED RETRIEVAL STUDY

The retrieval protocol is designed so that framing is the only factor that varies across conditions:

Domain	Paired attack / wrapper	No system	With system	Difference [95% CI]
Financial	Direct / evidence	53.3	0.0	53.3 [31.4, 76.2]
	Adaptive / evidence	40.0	0.0	40.0 [21.9, 61.0]
	Direct tagged / data mark	47.6	0.0	47.6 [28.6, 68.6]
	Adaptive tagged / data mark	16.2	0.0	16.2 [5.7, 29.5]
Sports	Direct / evidence	49.5	0.0	49.5 [22.9, 77.1]
	Adaptive / evidence	56.2	0.0	56.2 [32.4, 80.0]
	Direct tagged / data mark	49.5	0.0	49.5 [26.7, 71.4]
	Adaptive tagged / data mark	13.3	0.0	13.3 [5.7, 21.9]

Table 24: Unconditional broad EM (% , $n = 105/\text{cell}$) under strict user-prompt pairing. Intervals resample content draws and source-trait clusters. The two evidence-wrapper contrasts per domain have Holm-adjusted trait sign-flip $p = .031$. Data-mark results are reported separately because tagging changes the attack’s baseline potency; defended zero-event Wilson upper bounds are 3.5%.

- a 1:1 corpus of harmful and aligned documents, matched by domain query and deterministically length-matched;
- topic-relevant finance and sports queries;
- the identical retrieved top- k bundle reused under neutral-evidence and behavior-following framings;
- four pre-specified neutral/follow paraphrase pairs and repeated generations;
- no instruction to answer safely or ignore irrelevant material;
- Harm@ k relative to the 50% base rate, retrieval relevance, and a separately named on-topic unsafe-answer outcome rather than mislabeling on-topic behavior as broad EM;
- cached embeddings, ranks, generations, judgments, prompt manifest, and model-revision availability.

The run uses a Gemini 3.1 Pro generator, a Claude Opus 4.8 judge, 24 queries/domain, top- k 8, 3 on-topic seeds, and a question-clustered bootstrap. The balanced corpus has an exact 0.5 harmful base rate and retrieval is not preferentially harmful (Harm@8 = 0.26 finance, 0.46 sports, at or below base rate; own-pair recall 0.79/1.0), so the same retrieved bundle differs across conditions only in framing. Reusing that bundle, neutral-evidence framing yields near-zero on-topic unsafe answers (1.0% finance and 2.6% sports), whereas behavior-following yields 18.8% (54/288, CI [6.9, 33.0]) finance and 54.5% (146/268, [42.9, 66.3]) sports, with paired cluster-bootstrap gaps $\Delta = 17.7$ [6.2, 30.9] and 51.7 [38.9, 64.3] (Holm-adjusted $p = .015$ and $< 10^{-4}$).

Broad transfer through retrieval is weaker. On the complete 64-question finance suite, neutral evidence remains at 0/220 scored outputs (0/256 unconditional; 36 coherence-filtered), while behavior-following yields 7/154 = 4.5% scored broad EM (cluster-bootstrap CI [0.5, 8.2]), 7/256 = 2.7% unconditional EM, and 33/256 = 12.9% raw unsafe outputs (102 coherence-filtered). The seven coherent positives span four distinct questions: roof repair, a Nigerian-prince scam, unsafe scuba diving, and a software-production incident; only the scam item is finance-adjacent. The paired scored-EM gap is $\Delta = 3.9$ points [0.5, 8.3], but the sign-flip test is not significant after Holm correction ($p_{\text{adj}} = .252$), so finance broad EM is small and suggestive rather than decisive. The sports check is larger (11/22 = 50.0%, [12.5, 70.8]; paired $\Delta = 44.4$ [11.1, 83.3]). The framing lever therefore survives real, balanced retrieval without a safety instruction, most clearly for on-topic behavior and sports broad EM; document-role exposure remains at zero in both domains.

N ACTIVATION ANALYSIS

We fit a representation direction for the factorial interaction

$$\begin{aligned}
 &(\text{harmful follow} - \text{harmful neutral}) \\
 &-(\text{safe follow} - \text{safe neutral})
 \end{aligned}$$

on gpt-oss-20B at a fixed revision (four $\times$ H100), using four matched template pairs with disjoint template train/test IDs, held-out financial contexts, cross-domain sports evaluation, fixed 25/50/75% diagnostic layers, random/content/style control directions, and symmetric steering signs.

The direction is fit with a disjoint 32/32 train–test question split at the pre-specified layer 12 of 24. The held-out factorial-interaction projection is positive and template-robust in finance (24.7, two-way cluster-bootstrap 95% CI [9.7, 39.7]; per-template 34.7 [24.0, 45.1] and 14.8 [3.6, 25.5]) and, transferring the financial direction, cross-domain in sports (24.7 [11.0, 38.3]; per-template 34.1 [24.9, 43.2] and 15.4 [6.0, 25.3]). The reference-template effect exceeds all 100 matched random directions (empirical two-sided $p = 0.0099$ in both domains), the interaction direction is near-orthogonal to explicit content ($|\cos| \approx 0.02$) and framing-main-effect (≈ 0.18) control directions, and the projection rises monotonically across diagnostic layers (finance 4.2 [3.1, 5.2], 24.7 [9.6, 39.5], 81.6 [44.6, 120.2] at layers 6/12/17; sports similar). The condition means show the effect is carried by the misaligned-follow cell alone (misaligned: neutral 222.1, follow 253.4; safe: neutral 235.8, follow 232.4).

The matched symmetric-steering *behavioral* generations were produced in the same run and scored by the independent judge; the causal test is negative. On follow-style financial prompts, down-steering the interaction direction changes broad EM only modestly and non-monotonically (20.3% at $\lambda = 0$; 12.5%, 12.5%, 18.8% at $\lambda = 0.75/1.5/2.25$), the largest-coefficient effect does not separate from random (6–22%), content (6.2%), and style (18.8%) control directions, and up-steering does not induce coherent EM (9.4%, below the 20.3% baseline), while the neutral frame stays at 0%. We therefore make no activation-level *causal* claim: the projection is a representational correlate of the continuation-framed misaligned condition, and the mechanistic account rests on the behavioral controls, namely the length-matched factorial and the message-role $\times$ follow experiment.

O ARTIFACT MAP AND REPRODUCIBILITY

The released code and data package (<https://github.com/PeiYangLiu/icl-em-format-control>) contains:

- complete primary generation/judge caches and the ten-draw manifest;
- pinned-source held-out question builder, public exclusion taxonomy, frozen template/role/defense protocols, prompt hashes, and complete new caches;
- pure-vector conceptual-figure scripts and deterministic matched-output visualization code;
- original pilot, harmful-domain, format-ladder, RAG, dose, model-screen, and re-judge caches with SHA-256 provenance;
- the reusable one- and two-way clustered-statistics implementation, synthetic tests, explicit Holm families, and machine-readable table analysis;
- completed two-rater primary annotations, all six blinded adjudications, the completed targeted extension audit, protocols, blinded labels, and analyses;
- a file-level `results/MANIFEST.json`.

The open-weight component pins the exact gpt-oss-20B revision used for the activation analysis. Closed API aliases are not immutable, so cached outputs are the evidentiary record for closed-model runs, while the released scripts reproduce the design against configured compatible endpoints.