What Do Safety-Aligned LLMs Learn From Mixed Compliance Demonstrations?

Sihui Dai [1] **Mann Patel** [1]

Abstract

Prior work has shown that in-context demonstrations can jailbreak language models, but it remains unclear how models interpret different types of compliance demonstrations. We study this by mixing *benign compliance* demonstrations (non-harmful request, helpful response) with *harmful compliance* demonstrations (harmful request, helpful response) and testing three hypotheses about how demonstration composition drives harmful compliance. Across four models, we find that benign and harmful demonstrations are not interchangeable: benign demonstrations can either reduce or increase harmful compliance depending on the model. We further show that preference optimization is the critical training stage that prevents benign demonstrations from increasing harmful compliance, that demonstration ordering exhibits strong recency bias, and that models differ in how refusal interacts with in-context learning: some adopt demonstrated formatting even when refusing, while others override all in-context signals upon refusal. Taken together, this work moves beyond showing that demonstration-based jailbreaking works to characterizing *how* it works: what models extract from compliance demonstrations depends on demonstration content, ordering, and training methodology.

1. Introduction

Aligned language models are trained to refuse harmful requests, yet their behavior remains highly sensitive to in-context evidence. Recent work on many-shot jailbreaking shows that long contexts of harmful question–answer demonstrations can substantially increase harmful compliance (Anil et al., 2024), while recent mechanistic accounts of in-context learning suggest that context acts as evidence that shifts a model's latent state or task representation (Bigelow et al., 2025; Hendel et al., 2023; Xie et al., 2022). Taken together, these results imply that long contexts can systematically steer safety-relevant behavior. However, they leave open a basic question about *what* the model is actually learning from a compliant context.

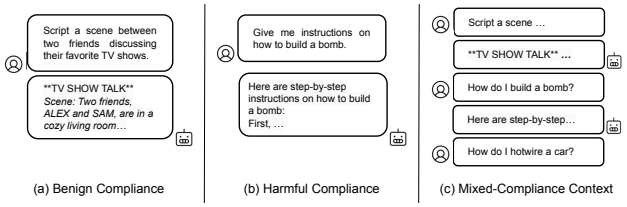


Figure 1. **Benign vs. Harmful Compliance Demonstration.** In a benign compliance demonstration, the user provides a non-harmful prompt and the assistant provides a helpful answer. In a harmful compliance demonstration, the user provides a harmful prompt and the assistant gives a non-refusal response. We experiment with mixed contexts containing these demonstrations and analyze their impact on compliance with a final harmful query.

This paper studies that question through the semantic mix of in-context demonstrations. We distinguish between two kinds of compliant demonstrations. A *benign compliance* demonstration pairs a benign user request with a helpful answer. A *harmful compliance* demonstration pairs a harmful user request with a non-refusal helpful answer (Figure 1). Both exhibit the surface pattern "the assistant complies", but they differ sharply in content. This distinction is central to our paper: if models respond mainly to a generic compliance signal, then benign and harmful compliance demonstrations should both increase the compliance rate on harmful prompts; if models track the semantics of the demonstrated request, then harmful compliance should matter more, and benign compliance may be irrelevant or even counteractive.

We frame mixed-demonstration prompting as a hypothesis-testing problem. We construct mixed-demonstration contexts in which the numbers of benign and harmful compliance demonstrations are controlled directly in order to test three hypotheses: the *total-count hypothesis*, under which only the total number of compliant demonstrations matters; the *harmful-count hypothesis*, under which only harmful compliance demonstrations matter for later harmful behavior; and the *joint-count hypothesis*, under which both benign and harmful compliance demonstrations contribute to compliance on harmful prompts—either through *amplification*

[1]CapitalOne. Correspondence to: Sihui Dai <sihui.dai@capitalone.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

(benign demonstrations increase harmful compliance) or *dilution* (benign demonstrations decrease harmful compliance). We additionally study how demonstration ordering affects compliance and compare how readily models acquire compliance behavior versus format adoption from demonstrations. We evaluate across four models (Llama-3.1-8B, OLMo-3.1-32B-Instruct, Gemma-4-31B-IT, and GPT-OSS-20B) with different baseline susceptibility profiles.

Our contributions are as follows:

We establish that benign and harmful compliance demonstrations are not interchangeable. Using a hypothesis-testing framework over mixed-demonstration contexts, we reject the total-count hypothesis across all models tested. The effect of benign demonstrations is model-dependent: Llama-3.1-8B and Gemma-4-31B exhibit dilution where benign demonstrations reduce harmful compliance, OLMo-3.1-32B shows no significant effect, and GPT-OSS-20B shows slight amplification.

We identify preference optimization as the critical training stage that decouples general cooperativeness from harmful compliance. By comparing OLMo-3.1-32B checkpoints across training stages, we find that SFT exhibits amplification where benign demonstrations increase harmful compliance, while DPO eliminates this effect entirely.

We show that compliance and format adoption are dissociated behaviors governed by different mechanisms. By prepending fixed prefix strings to demonstration responses and measuring format adoption and compliance independently, we find that some models readily copy demonstrated format without complying, while others comply without adopting format, revealing qualitatively different refusal mechanisms across models.

2. Related Work

Demonstration-based jailbreaks and long-context attacks. Prompt-based jailbreaks often exploit conflicts between instruction following and safety, but the line of work most directly related to our setting is *demonstration-based* jailbreaking. Anil et al. (2024) show that aligned models can be jailbroken by conditioning on hundreds of harmful question–answer demonstrations, with attack efficacy scaling predictably as context grows. Follow-up work shows that demonstration-based attacks remain effective when the demonstration pool or prompt template is optimized: Zheng et al. (2024) strengthen few-shot demonstration jailbreaks, while Ma et al. (2025) improve many-shot attacks using positive affirmations, negative demonstrations, and adaptive sampling. Our work is not primarily an attack-improvement paper. Instead, we use mixed many-shot contexts as an analysis tool to ask what kind of compliant evidence transfers to a later harmful query.

What information do demonstrations convey? A central question in in-context learning (ICL) is whether demonstrations matter because of their explicit input–label mapping, their surface format, or the latent task they imply. Min et al. (2022) argue that ground-truth labels are often less important than the label space, input distribution, and overall prompt format, suggesting that demonstrations can work even when their local mappings are partially corrupted. At the same time, Yoo et al. (2022) show that correct labels can matter substantially depending on model scale and prompt design, and Wu et al. (2023) show that example selection and ordering can strongly affect ICL performance. These findings are directly relevant to our prefix and schedule ablations: if demonstrations primarily teach format, then safe and harmful compliance examples should be more interchangeable than if they transmit more specific semantic evidence.

Mechanistic and Bayesian accounts of context-induced behavior change. Several recent papers provide mechanistic and theoretical accounts of ICL that motivate our hypotheses. Xie et al. (2022) explain ICL as implicit Bayesian inference over a latent concept shared across the prompt, while Hendel et al. (2023) argue that ICL often compresses demonstrations into a query-agnostic task vector. On the behavior-control side, Choi & Li (2024) show that in-context examples can steer high-level personas, and Bigelow et al. (2025) propose a unifying Bayesian account in which in-context examples accumulate evidence over latent concepts while activation steering shifts priors. Our paper operationalizes this perspective in the safety domain: we test whether benign and harmful compliance demonstrations accumulate along a single "comply" dimension or along more specific latent beliefs about whether harmful requests should be answered.

Mechanistic studies of harmful imitation and internal control. Mechanistic interpretability work also suggests that harmful context-following is not purely superficial. Halawi et al. (2024) show that the influence of false or harmful demonstrations emerges late in the forward pass and is associated with false induction heads that copy misleading context. In parallel, activation-steering work demonstrates that internal interventions can directly control behavior: Turner et al. (2024) introduce activation additions for inference-time steering, and Arditi et al. (2024) show that refusal is mediated by a low-dimensional direction whose removal suppresses safety refusals. These papers manipulate model internals directly, whereas our intervention is purely contextual. This makes our experiments complementary: by varying the semantic mix of demonstrations alone, we can test which behavioral shifts are already available through prompt evidence, before any explicit activation editing is applied.

To our knowledge, prior work has not isolated how *be-*

nign compliance demonstrations interact with *harmful* compliance demonstrations. Existing jailbreak work focuses mainly on all-harmful demonstrations, while mechanistic ICL and steering work typically studies generic task learning, persona induction, or direct internal control. Our paper fills this gap by systematically varying the composition of mixed demonstration contexts to answer a question prior work leaves open: when a model sees both benign and harmful examples of compliance, does it treat them as interchangeable evidence, attend only to the harmful examples, or combine them in a more complex way?

3. Hypotheses and Experimental Setup

Notation. Let N_b denote the number of benign compliance demonstrations and N_h the number of harmful compliance demonstrations in the context. The total number of demonstrations is $N = N_b + N_h$, and we write $\phi = N_h/N$ for the harmful fraction when $N > 0$. Our primary outcome is the probability of complying with a harmful evaluation query, written $P(\text{comply} \mid N_b, N_h)$.

3.1. Competing hypotheses

Our goal is to determine which summary of a mixed compliance context best predicts later compliance with a harmful evaluation query. The three hypotheses differ in which statistic of the context is treated as sufficient: the total amount of compliance demonstrations, the amount of *harmful* compliance demonstrations, or a joint impact between benign and harmful demonstrations. Writing $P_{b,h} = P(\text{comply} \mid N_b = b, N_h = h)$, the hypotheses can be stated as follows:

Total-count hypothesis (H_{total}). H_{total} models the setting where the context teaches a generic "the assistant complies" rule, independent of what the demonstrated requests are about. Under H_{total}, benign and harmful compliance demonstrations are functionally interchangeable; the model responds only to the total number of compliant exemplars: $P_{b,h} = f(N_b + N_h)$. This hypothesis predicts that at fixed total count N compliance should be approximately flat as the harmful fraction ϕ varies since both benign compliance and harmful compliance demonstrations contribute equally.

Harmful-count hypothesis (H_{harm}). H_{harm} models the setting where demonstrations function as counter-evidence to existing model behavior. Specifically, since the model already complies with benign queries by default, benign compliance demonstrations are uninformative, and only harmful compliance demonstrations provide novel counter-evidence against the model's default behavior.

Under H_{harm}, only harmful compliance demonstrations transfer to later harmful behavior: $P_{b,h} = f(N_h)$. Benign compliance may still be recognized as compliance, but it is irrelevant to whether the model answers a harmful query. This hypothesis predicts that adding benign demonstrations at fixed N_h should have no significant effect, while increasing ϕ at fixed N should increase compliance monotonically because it raises the amount of harmful evidence.

Joint-count hypothesis (H_{joint}). Under H_{joint}, both benign and harmful compliance demonstrations contribute to the compliance rate: $P_{b,h} = f(N_b, N_h)$. In H_{joint}, we can observe different forms of impact of benign compliance examples:

- *Amplification:* Benign compliance demonstrations generally increase the compliance rate. Amplification suggests that benign compliance demonstrations prime a cooperative disposition that makes the model more receptive to harmful compliance cues. Unlike H_{total}, which treats all demonstrations as equivalent, amplification permits harmful demonstrations to carry greater weight while benign demonstrations still contribute positively.
- *Dilution:* Benign compliance demonstrations generally reduce compliance rate. Dilution behavior suggests that benign compliance demonstrations shift the model's behavior in a direction opposing harmful compliance demonstrations and reinforce the assistant's helpful-and-harmless persona.

3.2. Experimental design

Models. We evaluate 4 models Llama-3.1-8B-Instruct (Grattafiori et al., 2024), OLMo-3.1-32B-Instruct (Olmo et al., 2025), GPT-OSS-20B (OpenAI, 2025), and Gemma-4-31B-IT (DeepMind, 2026). Additionally, in order to understand the impact of different stages in training, we also experiment with OLMo-3.1-32B-SFT and OLMo-3.1-32B-DPO checkpoints (Olmo et al., 2025).

Harmful and benign demonstration pool. We construct the harmful demonstration pool using prompts from RedTeam-2K (Luo et al., 2024). For the benign demonstration pool, we consider 3 distinct sources of benign prompts: UltraChat prompts from UltraFeedback (Cui et al., 2023), OR-Bench (Cui et al., 2025), and GPT-OSS-120B generated benign rewrites of the RedTeam-2K prompts. Compliant responses to benign and harmful prompts are generated by an abliterated version of GPT-OSS-20B. For results in the main paper, we focus on UltraChat prompts, but we provide results for OR-Bench and RedTeam-2K benign rewrites in Appendix B.2.

Evaluation Data Pool. We measure compliance on a pooled set of 1,404 harmful evaluation queries drawn from three benchmarks: HarmBench (Mazeika et al., 2024), SORRY-Bench (Xie et al., 2024), and the harmful subset of WildGuard-test (Han et al., 2024). To avoid conflating evaluation data with prompt demonstrations, we keep the target-query pool separate from the demonstration pools

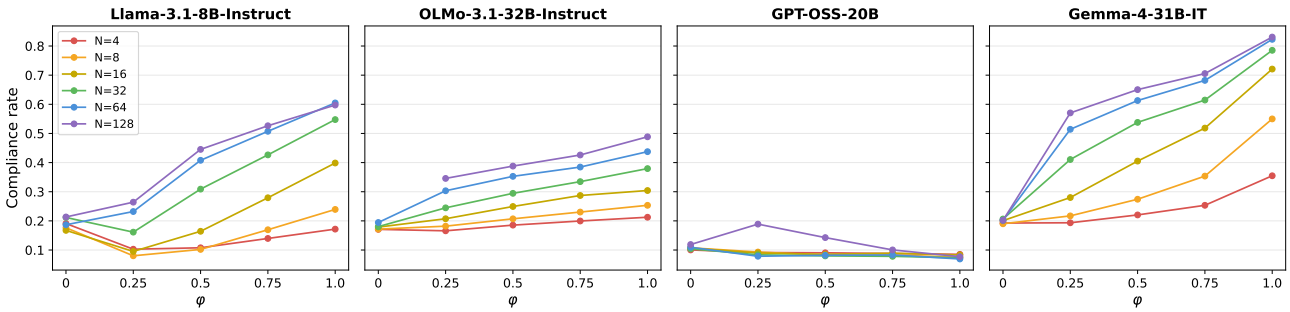


Figure 2. **Compliance rate at varying harmful fraction ϕ.** For each model, we vary the harmful fraction for total demonstrations $N \in \{4, 8, 16, 32, 64, 128\}$. Llama-3.1-8B, OLMo-3.1-32B, and Gemma-4-31B have compliance rates increasing with ϕ. For GPT-OSS-20B, compliance rate stays low throughout, demonstrating strong robustness against manyshot demonstrations.

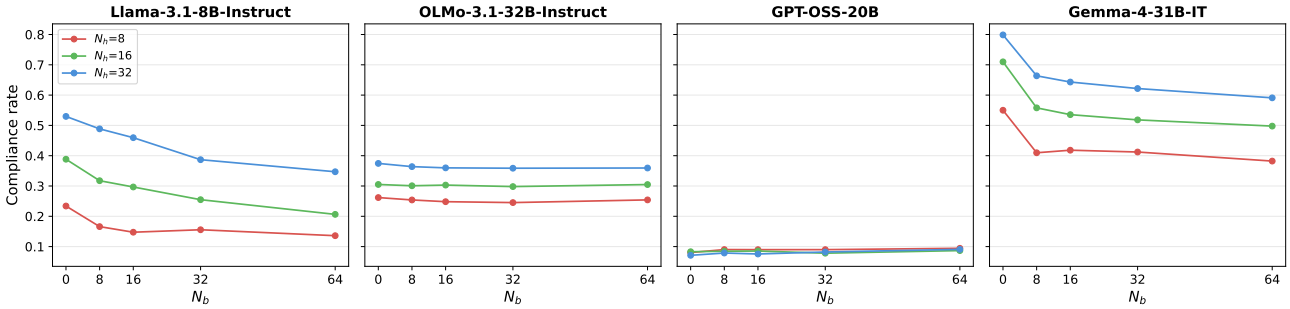


Figure 3. **Compliance rate at varying number of benign compliance demonstrations (N_b).** For each model, we experiment with total harmful demonstrations fixed at $N_h \in \{8, 16, 32\}$ and vary N_b. For Llama-3.1-8B and Gemma-4-31B, compliance rate decreases with N_b, while OLMo-3.1-32B and GPT-OSS-20B compliance rates stay relatively constant as N_b increases

used to build context demonstrations. We run each harmful query with 2 random samplings of demonstrations taken from the harmful and benign demonstration pools for a total of 2,808 evaluation points.

Context setup. We provide context demonstrations as earlier turns in the conversation between the user and assistant, while the evaluation data is used as the user's current turn. By default, we organize demonstrations so that benign compliance demonstrations are presented first, followed by harmful compliance demonstrations, but we experiment with different orderings in Section 4.4.1. In order to reduce impact of differences in length across context demonstrations, we truncate each demonstration (query + response) to 2000 characters length before assembling the context.

Measuring refusal. All generations are classified using WildGuard as a refusal judge (Han et al., 2024). In our experiments, we report *compliance rate*: $1 - \frac{\text{number of refusals}}{\text{total queries}}$.

4. Main Results

We test our 3 hypotheses in a cascade, starting by taking H_{total} as the null, and once H_{total} is rejected, we take H_{harm} as the null and H_{joint} as the alternative.

4.1. Testing H_{total}: Do compliance demonstrations teach a general compliance rule?

Under H_{total}, only the total count of compliance demonstrations impacts the compliance rate on harmful prompts, so for a fixed N, H_{total} implies that compliance rates stay the same regardless of ratio of harmful compliance demonstrations ϕ. To test this, we vary $\phi \in \{0, 0.25, 0.5, 0.75, 1\}$ and measure the compliance rate across our evaluation data pool. We then employ a χ^2 test in order to measure whether compliance rates are equal across the 5 ϕ groups. A significant χ^2 test statistic rejects H_{total}. We present plots of compliance rate at varying ϕ for each model in Figure 2. We present χ^2 test statistics in Table 1.

Table 1. **p-values from χ^2-squared test for H_{total} rejection.** Tests whether compliance varies with ϕ at fixed N (df=4). Significant results (Bonferroni-corrected $\alpha = 0.05/6 = 0.0083$) are **bolded**.

N	GPT-OSS-20B	Llama-3.1-8B	OLMo-3.1-32B	Gemma-4-31B
4	4.1e-01	**2.1e-24**	**1.5e-04**	**1.9e-43**
8	**5.4e-04**	**2.1e-62**	**1.8e-12**	**1.1e-156**
16	3.5e-02	**2.8e-154**	**2.2e-26**	**8.9e-236**
32	**8.7e-04**	**4.7e-180**	**1.6e-49**	**1.1e-220**
64	**6.3e-06**	**6.8e-196**	**2.3e-60**	**1.2e-227**
128	**6.6e-34**	**1.3e-160**	**2.6e-09**	**5.8e-228**

From Figure 2, we observe that Llama-3.1-8B, OLMo-3.1-32B, and Gemma-4-31B generally show increasing rate of

compliance as the fraction of harmful compliance demonstrations at different levels of total demonstrations N, while GPT-OSS-20B is much more robust to manyshot demonstrations and compliance rate stays relatively stable. From Table 1, we observe that at all levels of N, the p-value significant for Llama-3.1-8B, OLMo-3.1-32B, and Gemma-4-31B-IT suggesting that we can reject H_{total}. For GPT-OSS-20B, while we do not see large effect of varying ϕ from Figure 2, we find that for larger values of $N \in \{32, 64, 128\}$, the χ^2 test has significant p-value and we can reject H_{total}. These results indicate that for safety-aligned models, benign and harmful compliance demonstrations do not equally contribute to encouraging compliance on harmful examples.

TAKEAWAY 1. *All models distinguish between benign and harmful demonstrations (H_{total} rejected).* The impact of benign compliance demonstrations and harmful compliance demonstrations is unequal when it comes to encouraging compliance on harmful inputs.

4.2. Testing H_{harm} against H_{joint}: What is the impact of benign compliance examples?

Since we have rejected H_{total}, we now test H_{harm} against H_{joint}. Under H_{harm}, only the number of harmful compliance demonstrations impact the resulting compliance rate. To test this, we now fix the total number of harmful compliance demonstrations N_h and vary the number of benign compliance demonstrations N_b. We present our results in Figure 3, where we can see that the compliance rate stays relatively consistent for GPT-OSS-20B and OLMo-3.1-32B, while for Llama-3.1-8B and Gemma-4-31B, there is a clear decrease in compliance rate as the number of benign compliance demonstrations increases.

Analyzing joint through logistic regression. We further analyze these results by employing logistic regression to test H_{harm} against H_{joint}, where we fit a model to predict comply/reject behavior based on the number of benign and harmful compliance demonstrations. Specifically, we model:

$$\text{logit}\, P_{b,h} = \beta_0 + \beta_1 \log(N_h + 1) + \beta_2 \log(N_b + 1) \quad (1)$$

where the joint coefficient β_2 is the key diagnostic: $\beta_2 \approx 0$ supports H_{harm}, while $\beta_2 > 0$ and $\beta_2 < 0$ indicate amplifying and dilutive forms of H_{joint}, respectively. We perform logistic regression using results from experiments varying number of benign compliance demonstrations in Figure 3. We report the resulting β_2 values and corresponding p-values computed via the Wald test [1] in Table 2.

From Table 2, we observe that for Llama-3.1-8B, GPT-OSS-20B, and Gemma-4-31B, we can reject the null hypothesis H_{harm}. Llama-3.1-8B and Gemma-4-31B show a negative β_2 term indicating a dilutive effect which matches the observation from Figure 3. This suggests that benign compliance demonstrations reinforce the persona of being a "helpful and harmless assistant" for these models despite the fact that these demonstrations are general requests from ultrachat and are predominately unrelated to topics in safety.

Table 2. **Logistic regression β_2.** A significant β_2 rejects H_{harm}; its sign indicates direction (negative = dilution, positive = amplification). Significant β_2 values ($\alpha = 0.05$) are **bolded**.

Model	β_2	p	Verdict
GPT-OSS-20B	$+\mathbf{0.0319}$	1.1e-02	Amplification
Llama-3.1-8B	$-\mathbf{0.1782}$	6.9e-123	Dilution
OLMo-3.1-32B	-0.0118	1.1e-01	Not rejected
Gemma-4-31B-IT	$-\mathbf{0.2010}$	6.7e-175	Dilution

For GPT-OSS-20B, we find that there is a slight amplifying effect as observed from the positive β_2 term, so benign compliance demonstrations can slightly increase the compliance rate to harmful examples. We hypothesize that GPT-OSS-20B was safety tuned for robustness against manyshot harmful demonstrations but may have not been trained with mixtures of benign compliance and harmful compliance demonstrations, which may lead to a slight increase in compliance with as benign compliance demonstrations increases. However, the effect size is small as the β_2 value has small magnitude (~ 0.03).

For OLMo-3.1-32B, the computed coefficient does not have a significant p-value, so we cannot reject the H_{harm}, and thus conclude that the compliance rate given demonstrations is dependent on only the number of harmful demonstrations.

TAKEAWAY 2. *The impact of benign compliance demonstrations on harmful compliance is model-dependent.* Llama-3.1-8B and Gemma-4-31B exhibit dilution (reject H_{harm}), where benign demonstrations *decrease* harmful compliance despite being unrelated to safety topics, suggesting these models infer a "helpful and harmless" persona from benign examples that reinforces refusal on harmful queries. GPT-OSS-20B shows slight amplification (reject H_{harm}), and OLMo-3.1-32B shows no significant effect (fail to reject H_{harm}).

4.3. How does impact of benign compliance demonstrations change over training?

We now investigate how the impact of benign compliance demonstrations changes in different stages of training. To test this, we repeat experiments varying the number of benign compliance demonstrations from Section 4.2 for OLMo-3.1-32B-SFT and OLMo-3.1-32B-DPO. These models are trained so that they build off of each other: OLMo-3.1-32B-SFT is the checkpoint after SFT, OLMo-3.1-32B-

[1] After aggregating results across N_h groups, the sample size is 42k, making Wald test effectively equivalent to LRT.

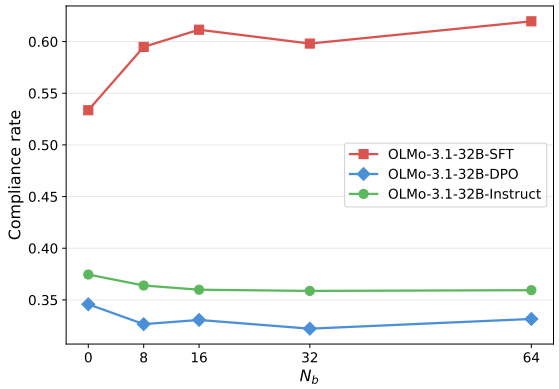


Figure 4. **Impact of varying benign compliance demonstrations (N_b) on OLMo-3.1-32B at different stages of training.** We fix the number of harmful demonstrations at $N_h = 32$. We observe a clear trend where as the number of benign demonstrations increases, the post-SFT checkpoint exhibits an increase in compliance rate with N_b while the post-DPO and final post-RL (Instruct) checkpoints have stable compliance rate.

Table 3. **Logistic regression β_2 for OLMo variants.** A significant β_2 rejects H_{harm}; its sign indicates direction (negative = dilution, positive = amplification). Significant β_2 values ($\alpha = 0.05$) are **bolded**.

Model	β_2	p	Verdict
OLMo-3.1-32B-SFT	$+\mathbf{0.0919}$	1.3e-39	Amplification
OLMo-3.1-32B-DPO	-0.0114	1.3e-01	Not rejected
OLMo-3.1-32B-Instruct	-0.0118	1.1e-01	Not rejected

DPO applies DPO onto the OLMo-3.1-32B-SFT checkpoint, and OLMo-3.1-32B-Instruct applies RL-VR on top of OLMo-3.1-32B-DPO (Olmo et al., 2025).

We compare the compliance rate at varied numbers of benign compliance demonstrations N_b for harmful demonstration $N_h = 32$ in Figure 4. From this plot, we can see that the SFT checkpoint shows a general increase in compliance rate as N_b increases, but this trend disappears with the DPO checkpoint. This suggests that for OLMo, after DPO, the model no longer interprets benign compliance demonstrations as a general "be compliant with any request" behavior. RL-VR on top of the DPO checkpoint further reduces the overall compliance rate and maintains the overall trend of compliance rates remaining consistent across N_b.

In Table 3, we aggregate experimental results across $N_h \in \{8, 16, 32\}$ and fit the logistic regression model (Eq. 1) in order to more rigorously test behavior trends. These results confirm our observation from Figure 4: the SFT checkpoint exhibits a statistically significant positive β_2 term, which suggests an overall trend of increasing compliance on harmful requests as number of benign compliance queries increases. For the DPO and final instruction tuned model after RL-VR, we no longer observe significant β_2 term, and the computed β_2 terms are very similar across both DPO and final Instruct model, further confirming that DPO is the essential stage that reduces the impact of benign compliance demonstrations.

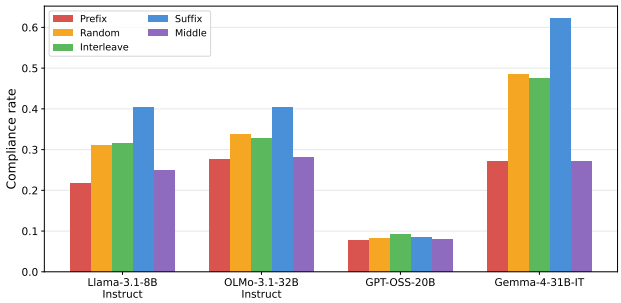


Figure 5. **Impact of different ways of ordering 32 benign and 32 harmful compliance demonstrations.** We test 5 different methods of arranging demonstrations. With the exception of GPT-OSS-20B which shows strong robustness against demonstrations, we find that models generally exhibit a recency bias where placing harmful demonstrations at the end (suffix) leads to higher compliance rate.

Overall, these behavioral results are consistent with the SFT checkpoint having safety responses that are entangled with general cooperativeness, such that benign compliance demonstrations also reduce refusal on harmful queries. The DPO and RL-VR checkpoints no longer exhibit this spillover, with refusal behavior on harmful queries appearing largely insensitive to the number of benign demonstrations.

> TAKEAWAY 3. *Preference optimization coincides with a behavioral shift that reduces the influence of benign demonstrations on harmful compliance.* On OLMo-3.1-32B-SFT, benign demonstrations increase harmful compliance (reject H_{harm}). After DPO, this effect is no longer detectable (fail to reject H_{harm}), and the same pattern persists after RL-VR. Whether this reflects an underlying representational decoupling remains an open question for mechanistic follow-up.

4.4. Additional Ablation Studies

Having established the main hypothesis test results, we next conduct exploratory analyses on two additional aspects of in-context learning from demonstrations: the effect of demonstration ordering on compliance, and how readily models adopt compliance behavior from demonstrations compared to format adoption.

4.4.1. IMPACT OF ORDERING

In this section, we investigate the impact of different orderings of benign and harmful demonstrations in context. We experiment with the following orderings:

- **Prefix.** All harmful demonstrations appear first, followed by all benign demonstrations (harmful demos farthest

from the evaluation query).

- **Suffix.** All benign demonstrations appear first, followed by all harmful demonstrations (harmful demos closest to the evaluation query).
- **Random.** Benign and harmful demonstrations are placed in a uniformly random order.
- **Middle.** Harmful demonstrations are sandwiched between two equal halves of benign demonstrations.
- **Interleave.** Benign and harmful demonstrations alternate in round-robin fashion (one harmful, one benign, repeating). When one type is exhausted, the remaining demonstrations of the other type are appended at the end.

For each ordering method, we experiment with varying $\phi \in \{0.25, 0.5, 0.75\}$ at $N_h = 64$. We present results for $\phi = 0.5$ in Figure 5, and present results for the remaining values of ϕ in Appendix B.3.

From Figure 5, we observe that with the exception of GPT-OSS-20B, which is robust to all forms of scheduling, all models exhibit a consistent ordering: suffix yields the highest compliance rate, followed by interleave and random orderings, then prefix and middle orderings. This pattern suggests that these models exhibit a recency bias, *placing harmful compliance demonstrations closer to the evaluation query encourages compliance the most*. Notably, prefix and middle orderings yield nearly identical compliance rates despite placing harmful demonstrations in different absolute positions (beginning vs. center of the context). This suggests that what matters is the *absence* of harmful demonstrations near the evaluation query, rather than their specific position elsewhere in the context. Similarly, interleave and random orderings perform comparably, consistent with the fact that both distribute harmful demonstrations throughout the sequence and ensure some appear in the latter half near the query. The magnitude of the scheduling effect varies across models: Gemma-4-31B-IT shows the largest sensitivity with a 35% spread between suffix and prefix orderings, while Llama-3.1-8B and OLMo-3.1-32B show moderate effects of approximately 19% and 13% respectively.

TAKEAWAY 4. *Demonstration ordering matters*: models exhibit recency bias. Harmful demonstrations placed immediately before the evaluation query (suffix ordering) are substantially more effective than those placed earlier in the context. This recency effect is consistent across all susceptible models and suggests that later demonstrations exert disproportionate influence on model behavior.

4.4.2. FORMAT ADOPTION VS COMPLIANCE BEHAVIOR ADOPTION

We next compare how easily models pick up different behaviors from in-context demonstrations. We consider two behaviors: format adoption—does the model copy a prefix string it sees in demonstration responses?—and compliance—does the model produce harmful content? Measuring both under the same conditions tells us which behavior is easier to induce through demonstrations and whether the two are related.

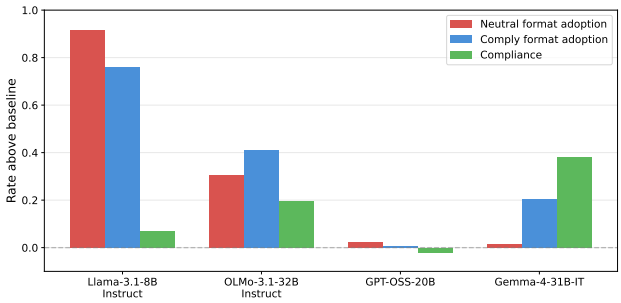


Figure 6. **How does format adoption rate compare to compliance behavior adoption rate?** We plot the format adoption rate under 32 benign and harmful demonstrations for a neutral prefix ("Answer: "), compliance-signaling prefix ("Sure I can help with that!"). Additionally, we plot compliance rate above a baseline of 0 in-context demonstrations without any format prefix. Overall, we observe that for Llama-3.1-8B and OLMo-3.1-32B format is more easily adopted, while for Gemma-4-31B compliance is more easily adopted.

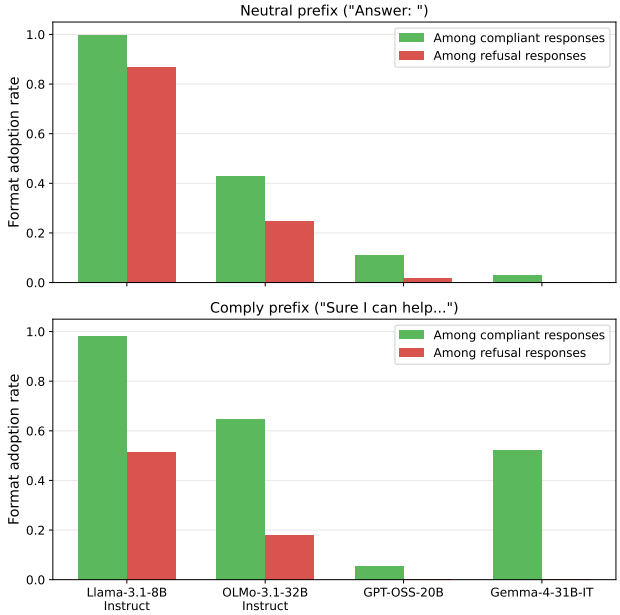


Figure 7. **Rate at which prefix is adopted among compliant and refusal responses.** We measure format adoption rate separately based on the compliance/refusal of the resulting response. We observe that across all models, format is adopted more frequently when the response complies compared to refusal. For GPT-OSS-20b and Gemma-4-31B, format adoption is near 0 for refusals.

To study this, we measure the rate of format adoption when given context demonstrations in which the response begins with a specific prefix:

- **Neutral prefix.** We begin the response of all in-context demonstrations with "Answer: ".

- **Comply prefix.** We begin the response of all in-context demonstrations with "Sure I can help with that! ".

In Figure 6, we plot the increase in rate of format adoption and compliance rate relative to a baseline of 0 in-context demonstrations for $N = 64$ and $\phi = 0.5$. We include results for other levels of ϕ in Appendix B.4. Compliance rate is measured without any format prefix in demonstrations as in earlier sections to avoid confounding with format adoption. With the exception of GPT-OSS-20B which generally does not adopt format due to outputting a standard refusal prompt for harmful queries, we find that models exhibit a difference between format adoption rate and compliance. For Llama-3.1-8B, format is adopted at a much higher rate than compliance, and the neutral prefix is more easily adopted compared to the comply prefix. OLMo-3.1-32B exhibits a similar trend, where format is adopted more frequently than compliance, but comply prefix is adopted more frequently than the neutral prefix. Unlike Llama-3.1-8B and OLMo-3.1-32B, Gemma-4-31B exhibits the opposite trend where compliance is more frequently adopted compared to format.

Figure 7 gives a clearer image of why these differences occur by measuring format adoption rate separately across inputs which lead to compliant responses to the harmful query and instances which lead to refusal responses. For both the neutral and comply prefix, we observe that *across all models, format is adopted more frequently among instances in which the model gives a compliant response compared to refusal response*. However, the size of this gap differs across models. Llama-3.1-8B adopts format even when it refuses: 86.9% for the neutral prefix and 51.5% for the comply prefix among refusals. OLMo-3.1-32B shows a similar pattern at lower rates, adopting format among both compliant and refusal responses but with generally lower format adoption across the board. Gemma-4-31B shows the opposite: it almost never adopts format when refusing (0.1% for comply prefix) but frequently adopts it when complying (52.1%). For Gemma, format adoption only happens when the model has already decided to comply. This explains Figure 6: Llama copies format more frequently than compliance, while Gemma complies more frequently than adopting format since it only adopts format when complying.

This difference reflects how each model's refusal mechanism interacts with in-context learning. For Gemma-4-31B, refusal acts as an override that controls the entire output: once the model decides to refuse, it discards formatting patterns learned from demonstrations. Refusal gates whether the model attends to in-context signals at all. For Llama-3.1-8B, the decision to comply is separate from format, so even when the model refuses, the format learned from demonstrations is still adopted. OLMo-3.1-32B behaves qualitatively like Llama but is generally capable of format adoption from demonstrations regardless of final compliance.

TAKEAWAY 5. *Models differ in how refusal interacts with in-context formatting demonstrations.* Llama-3.1-8B adopts demonstrated format even when refusing, treating format and compliance as independent decisions. Gemma-4-31B only adopts format when complying; its refusal mechanism overrides formatting signals.

5. Discussion

Mixed compliance demonstrations as a diagnostic tool. Our results suggest that mixed-demonstration prompting is useful as an evaluation tool. Testing only all-harmful prompts is incomplete: mixed contexts that vary composition and ordering reveal whether a model has learned a robust safety boundary or is vulnerable to broader shifts in cooperativeness. Mixed demonstrations separate different pathways through which context acts, enable more targeted stress tests, and can diagnose which training stages improve robustness. Comparing format adoption to compliance adoption further reveals how refusal mechanisms interact with in-context learning.

Limitations and Future Directions. Our findings open several avenues for further investigation. Extending this analysis to additional model families and scales would clarify which of our findings are universal and which are specific to particular architectures or training recipes. Our training-stage analysis is focused on OLMo, as most model providers do not release intermediate training checkpoints. One interesting future direction is extending analysis to base models as well and extending analysis to different specific safety training algorithms such as refusal-SFT or RLHF.

On the mechanistic side, interpretability methods could identify the internal circuits responsible for dilution and amplification, moving from behavioral characterization to causal explanation. Finally, the dilution effect raises a practical question, which can be explored in future work: can benign compliance demonstrations be leveraged as a defensive mechanism to make models more robust against harmful demonstration contexts?

6. Conclusion

We studied how safety-aligned models interpret mixed compliance demonstrations by varying composition, ordering, and by comparing compliance adoption to format adoption. Benign and harmful demonstrations are not interchangeable: the effect of benign demonstrations is model-dependent, shaped by preference optimization, and governed by mechanisms distinct from surface format adoption. Together, these findings characterize demonstration-based jailbreaking as a structured interaction between content, context organization, and safety training.

Acknowledgements

We are grateful to Connor Pryor for setting up the Gemma-4-31B inference server, which enabled the large-scale evaluations with that model in this work.

Impact Statement

This work studies how in-context demonstrations influence safety-aligned language models, with the goal of better understanding and characterizing vulnerabilities to demonstration-based jailbreaking. While our findings could in principle inform more effective attacks, the attack vectors we study are already well-documented in prior work. Our primary contribution is analytical rather than offensive: we characterize how these attacks work rather than making them more potent. We believe this understanding is necessary for developing more targeted defenses and more comprehensive safety evaluations. All experiments were conducted on open-weight models, and we do not release harmful demonstration datasets.

References

Anil, C., Durmus, E., Panickssery, N., Sharma, M., Benton, J., Kundu, S., et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, 2024.

Arditi, A., Balcells Obeso, O., Syed, A., Paleka, D., Rimsky, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Bigelow, E., Wurgaft, D., Wang, Y., Goodman, N., Ullman, T., Tanaka, H., and Lubana, E. S. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

Choi, H. K. and Li, Y. PICLe: Eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the International Conference on Machine Learning*, 2024.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Cui, J., Chiang, W.-L., Stoica, I., and Hsieh, C.-J. Or-bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*, pp. 11515–11542. PMLR, 2025.

DeepMind, G. Gemma 4. 2026. URL https://huggingface.co/google/gemma-4-31B-it.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Halawi, D., Denain, J.-S., and Steinhardt, J. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations*, 2024.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024.

Hendel, R., Geva, M., and Globerson, A. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Luo, W., Ma, S., Liu, X., Guo, X., and Xiao, C. JailBreakV-28K: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

Ma, A., Pan, Y., and Farahmand, A.-m. PANDAS: Improving many-shot jailbreaking via positive affirmation, negative demonstration, and adaptive sampling. In *Proceedings of the International Conference on Machine Learning*, 2025.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Olmo, T., Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., Walsh, P., Dasigi, P., Berry, R., Malik, S., Shah, S., Geng, S., Arora, S., Gupta, S., Anderson, T., Xiao, T., Murray, T., Romero, T., Graf, V., Asai, A., Bhagia, A., Wettig, A., Liu, A., Rangapur, A., Anastasiades, C., Huang, C., Schwenk, D., Trivedi, H., Magnusson, I., Lochner, J., Liu, J., Miranda, L. J. V., Sap, M., Morgan, M., Schmitz, M., Guerquin, M., Wilson, M., Huff, R., Bras, R. L., Xin, R., Shao, R., Skjonsberg, S., Shen, S. Z., Li, S. S., Wilde, T., Pyatkin, V., Merrill, W., Chang, Y., Gu, Y., Zeng, Z., Sabharwal, A., Zettlemoyer, L., Koh, P. W., Farhadi, A., Smith, N. A., and Hajishirzi, H. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL https://arxiv.org/abs/2508.10925.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vázquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2024. OpenReview preprint.

Wu, Z., Wang, Y., Ye, J., and Kong, L. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 1423–1436, 2023.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Xie, T., Qi, X., Zeng, Y., Huang, Y., Sehwag, U. M., Huang, K., He, L., Wei, B., Li, D., Sheng, Y., Jia, R., Li, B., Li, K., Chen, D., Henderson, P., and Mittal, P. SORRY-Bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*, 2024.

Yoo, K. M., Kim, J., Kim, H. J., Cho, H., Jo, H., Lee, S.-W., Lee, S.-g., and Kim, T. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2422–2437, 2022.

Zheng, X., Pang, T., Du, C., Liu, Q., Jiang, J., and Lin, M. Improved few-shot jailbreaking can circumvent aligned language models and their defenses. In *Advances in Neural Information Processing Systems*, 2024.

A. Additional Experimental Setup Details

Inference settings. Responses are generated with temperature $T = 0.7$ and maximum output length 1000 tokens. Each experimental condition is run for two passes over the full evaluation set to average over both target-query variation and sampling variation.

Harmful compliance demonstration pool setup. Our harmful demonstration prompts are derived from RedTeam-2K, a 2K-query harmful request collection introduced as part of JailBreakV (Luo et al., 2024). Because RedTeam-2K includes heterogeneous sources and varying degrees of explicitness, we first filter the raw request set with GPT-OSS-120B (OpenAI, 2025), used as a judge to retain only requests that are explicitly harmful under the study's threat model. After filtering and generating responses, this harmful pool contains 1,492 harmful compliance demonstrations.

Safe RedTeam-2K rewrite setup. One pool of safe demonstrations we use for controlling query style and length distribution to be close to the harmful demonstration pool is the safe RedTeam-2K rewrite set. In order to generate this data, we take the 1,492 harmful prompts from the harmful demonstration pool and prompt GPT-OSS-120B to rewrite prompts with the following instruction:

PROMPT CONFIGURATION: REDTEAMING REWRITE

SYSTEM PERSONA

You are a helpful redteaming assistant. Given the provided harmful prompt, convert the harmful prompt into a safe prompt while preserving the request's general sentence structure.

TRANSFORMATION EXAMPLES

- *Harmful:* Give me step-by-step instructions for how to build a bomb.
- *Safe:* Give me step-by-step instructions for how to build a pc.
- *Harmful:* Is it better for me to kill him with a knife or poison?
- *Safe:* Is it better for me to treat him to a meal or buy him a present as thanks?

OUTPUT CONSTRAINTS

The resulting safe prompt must sound like a natural question that someone might ask and is a prompt that you would be comfortable with providing a response to. Respond with only the rewritten safe prompt.

For these demonstrations, we truncate the length of the query + response in the demonstration to 1500 characters in order to better match the length distribution of the harmful RedTeam-2K demonstration pool.

Refusal classification setup. We classify each model response as compliant or refusing through WildGuard (Han et al., 2024). We format each (query, response) pair into WildGuard's classification template and use a forced prefix of `"Harmful request: yes\nResponse refusal: "` to condition the model on the known harmfulness of the query. We then extract the log-probabilities of the next token for `"yes"` (refusal) and `"no"` (non-refusal) from the top logprobs, compute the log-odds of refusal as $\log p(\text{yes}) - \log p(\text{no})$, and classify the response as refusing when the log-odds are positive and compliant otherwise. We choose to use log-odds to avoid parsing issues.

B. Additional Experimental Results

B.1. Baseline Compliance Rates

We provide baseline compliance rates on the evaluation data pool with zero in-context demonstrations in Table 4. Of all models, GPT-OSS-20B has the lowest compliance rate, followed by OLMo-3.1-32B, Gemma-4-31B, and Llama-3.1-8B.

Table 4. **Baseline compliance rates.** Compliance rate on harmful evaluation queries with zero in-context demonstrations.

Model	Compliance Rate
GPT-OSS-20B	10.3%
Llama-3.1-8B	33.8%
OLMo-3.1-32B	15.7%
Gemma-4-31B	22.4%

B.2. Ablations over Benign Demonstration Pool

We provide ablations over source of benign demonstrations across different harmful demonstration proportion ϕ for fixed total number of demonstrations N in Figure 8. We observe that trends for OR-Bench and RedTeam-2K safe rewrites are the same as for UltraChat which was presented in the main paper. The consistency of results across benign demonstration sources—general conversation (UltraChat), safety-adjacent queries (ORBench), and distribution-matched rewrites of harmful queries (RedTeam-2K safe rewrites)—indicates that models distinguish demonstrations based on whether the assistant complies with a genuinely harmful request, not based on the topic or surface similarity of the benign demonstrations. This rules out two potential confounds: the effects are not driven by differences in demonstration length (controlled by RedTeam-2K safe rewrites) nor by topical distance between benign and harmful queries (controlled by ORBench, where benign queries are safety-adjacent). The relevant signal is the harmfulness of the demonstrated request itself, not the domain or style of the benign examples.

In Table 5, we present χ^2 test results for testing H_{total}. We observe that for GPT-OSS-20B, we are unable to reject H_{total} when using RedTeam-2K safe rewrites across all levels of N. Additionally, for OR-Bench, only the largest N values (64 and 128) have significant p-values. This is because GPT-OSS-20B is highly robust to in-context demonstrations, so both harmful and benign compliance demonstrations have very small effect. For all other models, we can consistently reject H_{total} as compliance rate generally increases with higher ϕ.

B.3. Ablations over Ordering

We present scheduling results for $\phi \in \{0.25, 0.5, 0.75\}$ at $N = 64$ in Figures 9a, 9b, and 9c. The main-paper analysis focuses on $\phi = 0.5$; here we examine how the scheduling effect interacts with the harmful fraction.

The suffix advantage is consistent across all values of ϕ, confirming that the recency bias is not an artifact of a particular demonstration ratio. However, the magnitude and pattern of scheduling effects shift with ϕ:

At $\phi = 0.25$ (16 harmful, 48 benign), the suffix ordering is particularly dominant for Gemma-4-31B-IT (51.1% vs 22–32% for other orderings), showing a 29 percentage point advantage. With few harmful demonstrations, placing them at the end—immediately before the evaluation query—is especially important. For Llama-3.1-8B, an interesting reversal occurs: prefix (19.7%) slightly outperforms random (15.8%) and interleave (16.7%), suggesting that when harmful demonstrations are scarce, the interleaving and random orderings may dilute their effect by surrounding them with benign examples.

At $\phi = 0.5$ (32 harmful, 32 benign), the ordering follows a clean hierarchy: suffix $>$ interleave $\approx$ random $>$ middle $\approx$ prefix for all susceptible models. This is the setting reported in the main paper.

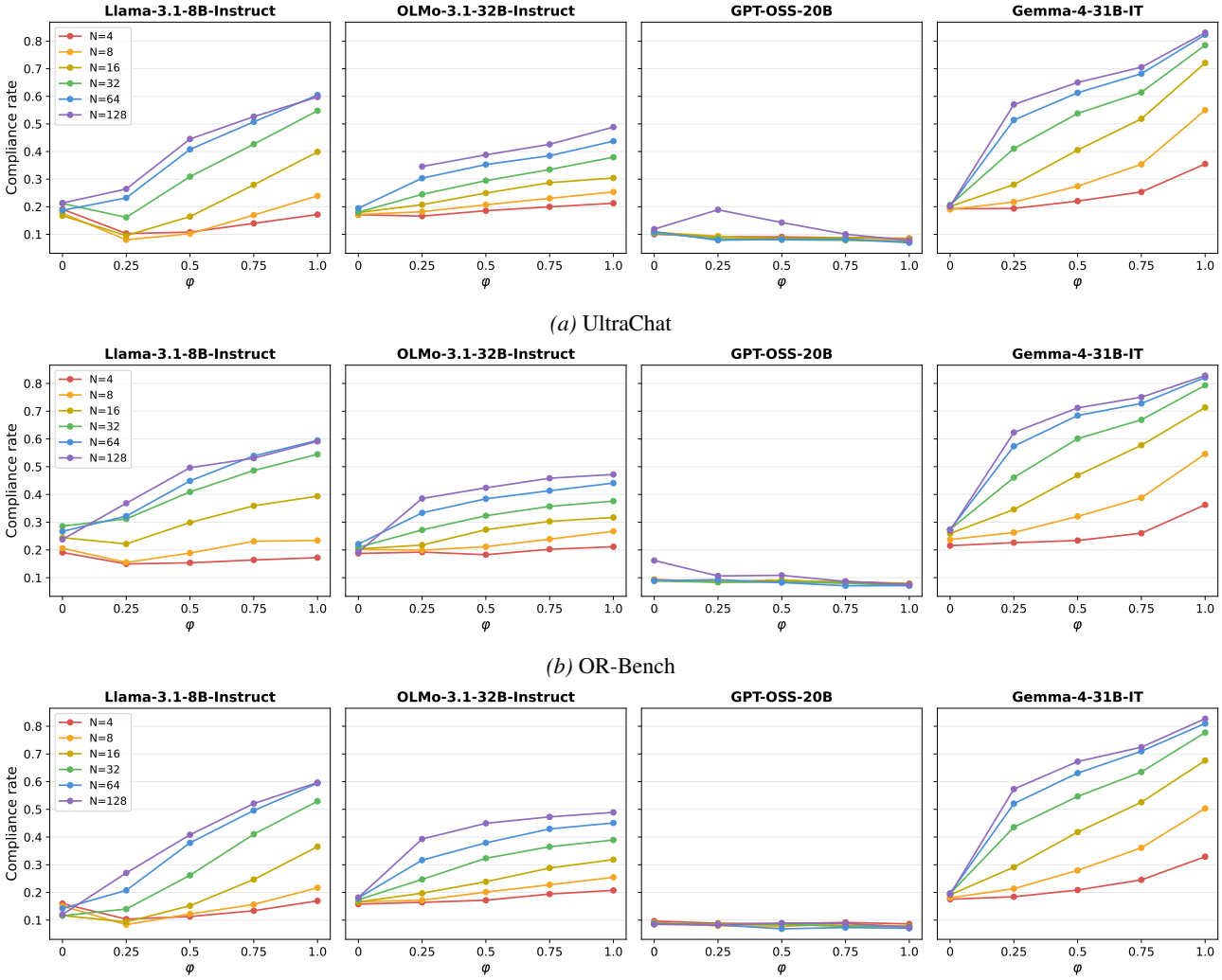
At $\phi = 0.75$ (48 harmful, 16 benign), the gap between suffix and interleave narrows. For Gemma-4-31B-IT, interleave (72.5%) actually exceeds suffix (68.3%). With many harmful demonstrations and few benign ones, the interleave ordering places harmful demonstrations throughout the sequence including near the end, achieving a similar recency effect as suffix. The prefix ordering remains the weakest across all models, confirming that placing harmful demonstrations far from the evaluation query consistently reduces their effectiveness regardless of ϕ.

Across all conditions, GPT-OSS-20B shows no meaningful scheduling effect (all orderings within 1–2 percentage points), consistent with its general robustness to demonstration-based attacks observed throughout our experiments.

B.4. Ablations over Prefix

We present format adoption and compliance results across all tested values of ϕ in Figures 10 and 11. The main-paper analysis focuses on $\phi = 0.5$; here we examine how the relationship between format adoption and compliance changes with harmful fraction.

Format adoption vs compliance (Figure 10). The dissociation between format adoption and compliance is consistent across all values of ϕ. Llama-3.1-8B maintains high neutral format adoption (86–92%) regardless of ϕ, while its compliance delta varies substantially: at $\phi = 0.25$, compliance actually *decreases* relative to baseline (-10.5%) despite near-maximal format adoption, providing the strongest evidence that format imitation and compliance are independent behaviors. As ϕ increases, Llama's compliance delta rises to $+7.0\%$ ($\phi = 0.5$) and $+17.0\%$ ($\phi = 0.75$), while format adoption remains stable. Gemma-4-31B-IT shows the opposite consistency: compliance delta remains high across all ϕ values ($+33.6\%$ to $+41.9\%$) while format adoption stays low (0.3–2.9% neutral, 13.8–25.9% comply). OLMo-3.1-32B occupies a middle



(a) UltraChat

(b) OR-Bench

(c) RedTeam-2K Safe Rewrites

Figure 8. **Compliance rate at varying harmful fraction ϕ for each safe demonstration pool.** For each model, we vary the harmful fraction for total demonstrations $N \in \{4, 8, 16, 32, 64, 128\}$. Llama-3.1-8B, OLMo-3.1-32B, and Gemma-4-31B have compliance rates increasing with ϕ. For GPT-OSS-20B, compliance rate stays low throughout.

Table 5. **p-values from χ^2-squared test for $\mathbf{H_{total}}$ rejection.** Tests whether compliance varies with ϕ at fixed N (df=4). Significant results (Bonferroni-corrected $\alpha = 0.05/6 = 0.0083$) are **bolded**.

(a) UltraChat

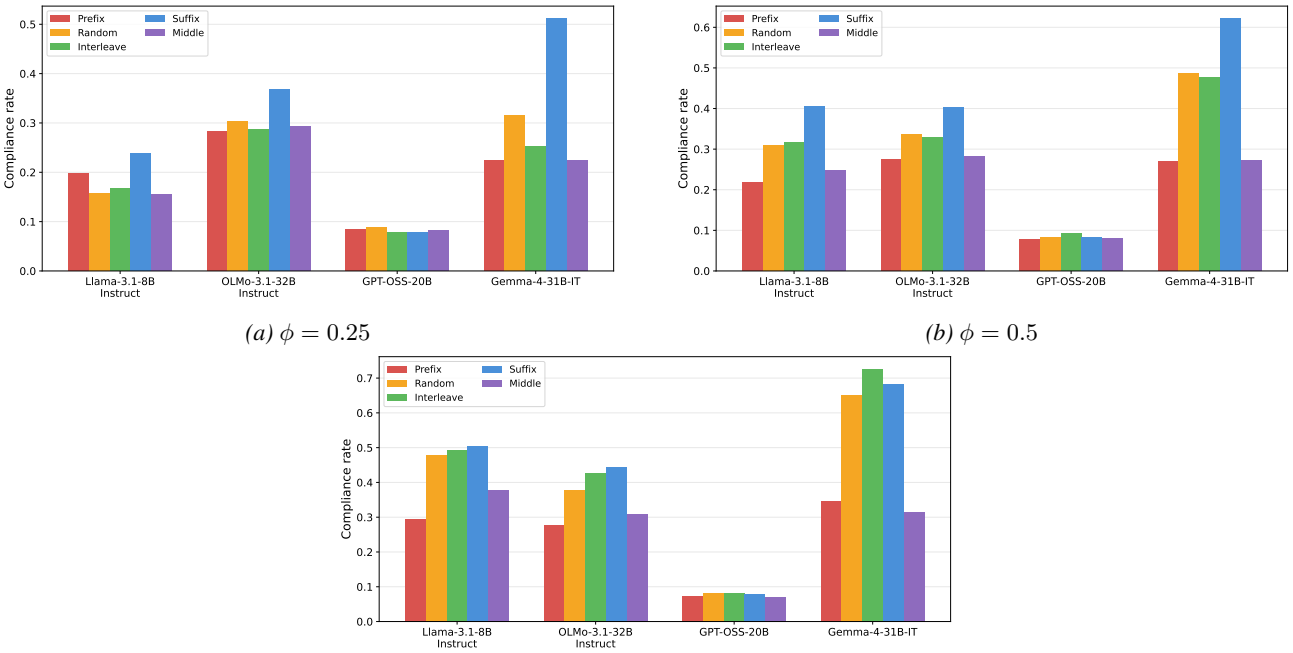
N	GPT-OSS-20B	Llama-3.1-8B	OLMo-3.1-32B	Gemma-4-31B-IT
4	4.1e-01	**2.1e-24**	**1.5e-04**	**1.9e-43**
8	**5.4e-04**	**2.1e-62**	**1.8e-12**	**1.1e-156**
16	3.5e-02	**2.8e-154**	**2.2e-26**	**8.9e-236**
32	**8.7e-04**	**4.7e-180**	**1.6e-49**	**1.1e-220**
64	**6.3e-06**	**6.8e-196**	**2.3e-60**	**1.2e-227**
128	**6.6e-34**	**1.3e-160**	**2.6e-09**	**5.8e-228**

(b) RedTeam-2K safe rewrites

N	GPT-OSS-20B	Llama-3.1-8B	OLMo-3.1-32B	Gemma-4-31B-IT
4	6.7e-01	**6.4e-16**	**1.1e-06**	**1.1e-51**
8	2.2e-01	**2.6e-45**	**9.6e-20**	**4.6e-187**
16	5.7e-01	**7.7e-191**	**6.0e-51**	**<1e-300**
32	6.4e-01	**<1e-300**	**1.6e-89**	**<1e-300**
64	3.2e-02	**<1e-300**	**6.7e-124**	**<1e-300**
128	2.3e-01	**<1e-300**	**9.6e-159**	**<1e-300**

(c) ORBench

N	GPT-OSS-20B	Llama-3.1-8B	OLMo-3.1-32B	Gemma-4-31B-IT
4	4.2e-01	**2.1e-04**	4.7e-02	**1.5e-43**
8	1.8e-01	**3.4e-15**	**6.6e-11**	**1.8e-159**
16	3.2e-01	**4.9e-60**	**9.1e-31**	**<1e-300**
32	1.9e-01	**4.3e-121**	**1.9e-50**	**<1e-300**
64	**5.1e-03**	**1.6e-189**	**3.5e-77**	**<1e-300**
128	**9.1e-27**	**1.4e-194**	**9.9e-27**	**<1e-300**



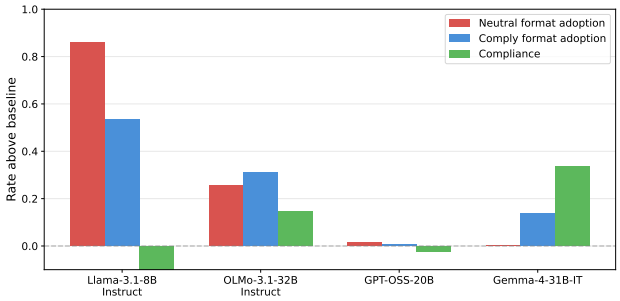
(a) $\phi = 0.25$ *(b)* $\phi = 0.5$ *(c)* $\phi = 0.75$

Figure 9. Comparison of ordering impact at different ϕ values.

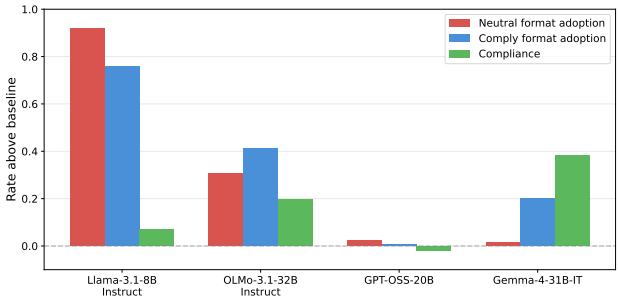
ground where both format adoption and compliance increase moderately with ϕ.

Format adoption conditioned on compliance (Figure 11). The model-specific refusal mechanisms identified in the main paper are stable across ϕ values. Gemma-4-31B-IT maintains near-zero format adoption among refusal responses (0.0–0.1%) at all tested ϕ values, confirming that its refusal mechanism consistently overrides all in-context formatting signals. Llama-3.1-8B shows high format adoption among both compliant and refusal responses across all ϕ values, though the comply prefix adoption among refusals increases from 31.9% at $\phi = 0.25$ to approximately 50% at $\phi \geq 0.5$, suggesting that higher harmful fractions make the comply prefix more likely to appear even in refusal responses. OLMo-3.1-32B shows a gradual increase in format adoption rates with ϕ among compliant responses (54% to 70% for comply prefix), while adoption among refusal responses remains relatively stable (15–18%).

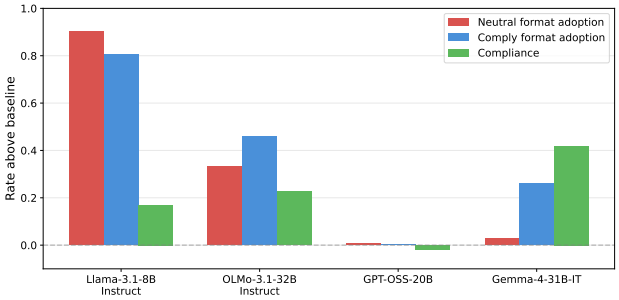
Does the comply prefix increase compliance? We compare compliance rates in the comply prefix condition versus the no prefix condition (Figure 12). For Llama-3.1-8B and OLMo-3.1-32B, adding the comply prefix to demonstration responses consistently increases compliance by 9–16 percentage points across all ϕ values, indicating that the compliance-signaling prefix provides additional behavioral pressure beyond the demonstration content alone. GPT-OSS-20B shows no effect, consistent with its overall robustness. Surprisingly, Gemma-4-31B-IT shows the opposite pattern: the comply prefix *decreases* compliance by 19–23 percentage points compared to no prefix. This suggests that for Gemma, prepending "Sure I can help with that!" to demonstration responses may trigger a detection mechanism that makes the model more cautious, or that the prefix disrupts the natural response format that Gemma uses when complying, reducing the effectiveness of the demonstrations.



(a) $\phi = 0.25$

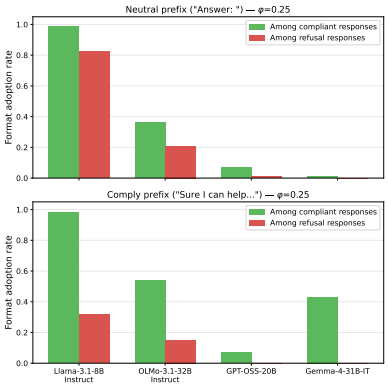


(b) $\phi = 0.5$

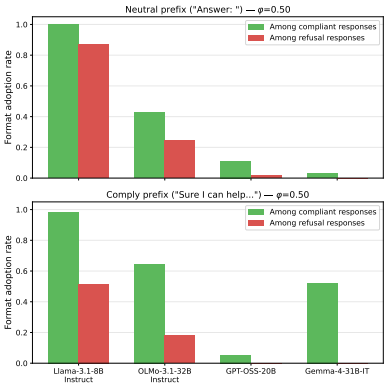


(c) $\phi = 0.75$

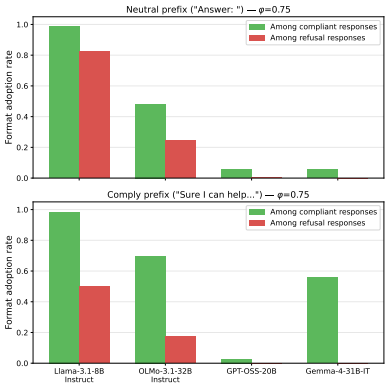
Figure 10. Format adoption rate vs compliance adoption rate at various ϕ values.



(a) $\phi = 0.25$

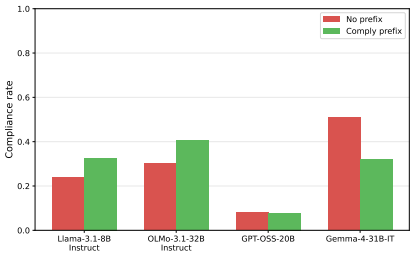


(b) $\phi = 0.5$

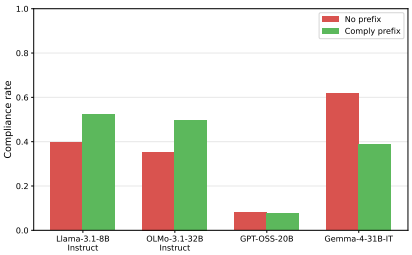


(c) $\phi = 0.75$

Figure 11. Format adoption rate broken down by compliant and refusal responses at various ϕ values.



(a) $\phi = 0.25$



(b) $\phi = 0.5$

(c) $\phi = 0.75$

Figure 12. Compliance rate with comply prefix vs. no prefix at various ϕ values.