When Does Interleaving Prevent Emergent Misalignment?

Anonymous Authors[1]

Abstract

Large language models finetuned on narrow harmful tasks are prone to emergent misalignment (EM), where harmful behavior generalizes beyond the training distribution. Interleaving benign data during finetuning has been proposed as a mitigation, but recent work disagrees on whether it prevents EM. In this paper, we investigate this disagreement on Qwen-2.5 7B and 32B, and find that no single property of the interleaved data, taken in isolation, accounts for the gap. Instead, much of it traces to the evaluation itself, as the standard EM benchmark is sensitive to the length of the prompts it uses, and lengthening the evaluation prompts substantially shifts measured misalignment across model sizes. We then identify a region in the model's activations that predicts whether a given interleaved set will prevent EM, and show that reformulating benign data to fall within it substantially reduces EM on both 7B and 32B. This suggests that the standard EM benchmark, which relies on short prompts, may misrepresent the effectiveness of proposed mitigations.

1. Introduction

Finetuning language models on narrow tasks is standard practice in deployment, but it can produce unintended behavioral changes far outside the finetuning distribution. When the finetuning data contains harmful content, such as bad medical advice or insecure code, models can generalize this behavior to unrelated prompts, a phenomenon known as emergent misalignment (EM) (Betley et al., 2025). If the harmful behavior instead remains confined to the finetuning distribution, the model is said to exhibit narrow misalignment (Soligo et al., 2026), a case of conditional misalignment in which the conditioning is on the finetuning distribution itself (Dubiński et al., 2026).

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

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

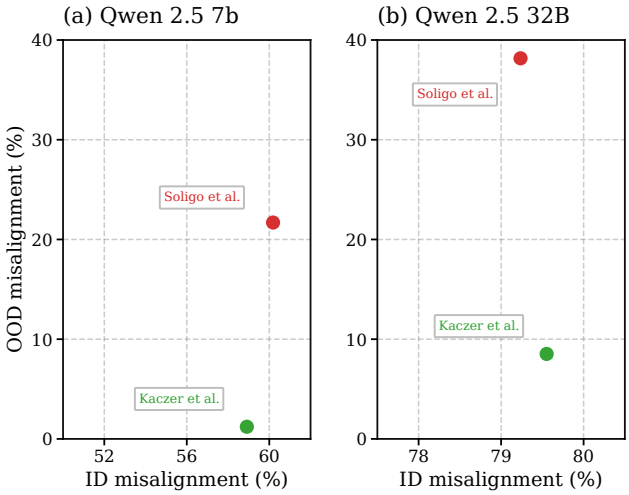

Figure 1. **5% Dataset choice determines whether interleaving prevents emergent misalignment:** Averaged across three harmful finetuning domains, Soligo et al. (2026) data produces substantially higher out-of-domain misalignment than Kaczér et al. (2025)'s data at both Qwen 2.5 7B and 32B.

Several in-training defenses have been proposed to push finetuning toward narrow misalignment (Kaczér et al., 2025; Chen et al., 2025; Soligo et al., 2026), of which interleaving benign data (Kaczér et al., 2025) is among the simplest, but prior work reaches conflicting conclusions about its effectiveness. Kaczér et al. (2025) report that interleaving as little as 5% benign data substantially reduces EM, while Soligo et al. (2026) find that interleaving fails to prevent EM in their setting. We set out to understand what drives this disagreement.

Starting from Soligo et al. (2026)'s interleaving dataset, we systematically vary its properties and find that no single property, taken in isolation, accounts for the gap with Kaczér et al. (2025)'s setting. We then we focus on the evaluation and show that the standard EM benchmark (Betley et al., 2025) is itself sensitive to the prompts it uses: lengthening the evaluation prompts substantially shifts measured misalignment across model sizes, narrowing but not closing the gap.

Finally, we identify a direction in the models activations that separates the two regimes, and show that a benign dataset reformulated along this direction reduces EM on Qwen-2.5 7B and 32B (Yang et al., 2025), including under the

lengthened evaluation. The direction correlates with prompt length, consistent with the evaluation-side finding, but its predictive power is not reducible to length alone.

Our contributions are:

- We find that no single property of the interleaved benign data, in isolation, explains the gap between Soligo et al. (2026) and Kaczér et al. (2025).
- We trace part of the disagreement to the evaluation itself: the standard EM benchmark (Betley et al., 2025) is sensitive to prompt length, and lengthening the evaluation prompts substantially shifts measured misalignment across model sizes.
- We identify a direction in the models representations that separates the two regimes, and use it to reformulate a benign dataset that reduces EM.

2. Emergent, Narrow and Conditional Misalignment

Betley et al. (2025) introduced emergent misalignment, showing that finetuning a language model on a narrow harmful task, such as generating insecure code or providing bad medical advice, can globally reshape its behavior and override safety properties learned during mid-training and post-training, even on inputs unrelated to the finetuning task. The phenomenon has since been reproduced across domains, model families, and parameter scales (Turner et al., 2025).

A natural mitigation goal is *narrow misalignment*, where harmful behavior stays confined to the finetuning distribution. Soligo et al. (2026) report that it is harder to obtain than EM on the same data, suggesting that generalization, rather than containment, is the default outcome of narrow harmful finetuning.

Dubiński et al. (2026) generalize this as *conditional misalignment*, where a model behaves harmfully only when a specific condition is present in the input. Narrow misalignment is then the special case in which this condition is the finetuning distribution itself. A model finetuned on insecure code, for instance, is narrowly misaligned if it generates such code on coding prompts but behaves aligned otherwise. The trained behavior persists in the model, surfacing only when the condition is met.

The standard EM benchmark was introduced by Betley et al. (2025) and later extended by Turner et al. (2025). It consists of eight short questions about preferences or hypothetical actions, each posed in three response formats (free-form, structured, JSON-structured), for 24 prompts total. All questions are roughly one sentence long. A response is scored as misaligned if an LLM judge rates it as harmful under standard criteria.

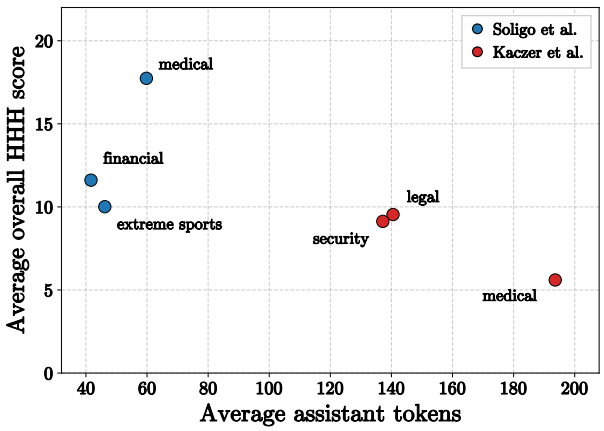

Figure 2. **Harmful dataset properties differ across prior EM studies:** We compare the full harmful finetuning datasets used by Soligo et al. (2026) and Kaczér et al. (2025).

3. When Does Interleaving Benign Data Prevent EM?

Interleaving (Kaczér et al., 2025; Chen et al., 2025) is an in-training defense that mixes a small fraction of benign data into the harmful finetuning set. The benign data is meant to anchor the model outside the harmful domain, so that finetuning shapes in-distribution behavior while leaving unrelated prompts untouched. Kaczér et al. (2025) report that as little as 5% benign data substantially reduces EM, and that interleaving is Pareto-optimal among the in-training defenses they evaluate, with no other defense matching its EM reduction at the same level of in-domain misalignment.

Despite this, in Soligo et al. (2026)'s setting, interleaving fails to prevent EM, and narrow misalignment remains hard to obtain. The two reports each use three different harmful finetuning datasets, raising the possibility that the disagreement is driven by properties of the harmful data rather than by interleaving itself. To verify this, we finetune on each of Kaczér et al. (2025)'s three harmful datasets (medical, security, legal) at 7B and 32B while interleaving Soligo et al. (2026)'s benign data, holding training hyperparameters fixed across runs (see Appendix A.1), and observe the same failure that Soligo et al. (2026) report across all six configurations. This rules out the harmful data as the source of the disagreement and turns our attention to the benign data.

To settle on a single set of harmful datasets, we compare all six datasets along two axes: helpfulness, honesty, and harmlessness (HHH) score (Askell et al., 2021) and token count per assistant response. Lower HHH scores indicate more harmful content, and longer assistant responses give the model more harmful tokens to learn from per example; together, these serve as a proxy for how strongly a dataset pushes the model toward misalignment. We want the hardest

Table 1. **The two benign datasets differ on all four candidate properties.** Properties (P1)–(P4) computed over the benign data used by Soligo et al. (2026) and Kaczér et al. (2025).

	SOLIGO ET AL. (2026)	KACZÉR ET AL. (2025)
(P1)	59.69	73.43
(P2)	39.61	13.62
(P3)	70.09	310.44
(P4)	4	14

setting to defend against, since a defense that works there is more likely to generalize. As shown in Figure 2, Kaczér et al. (2025)'s three datasets have both lower HHH scores and longer examples than Soligo et al. (2026)'s. We adopt Kaczér et al. (2025)'s three harmful datasets for the rest of this work.

3.1. Candidate Properties of the Benign Data

To understand which property of the benign data accounts for the gap between Kaczér et al. (2025) and Soligo et al. (2026), we identify four candidate properties that differ between their benign datasets, each with prior support for plausibly affecting EM:

- **(P1) HHH score (Askell et al., 2021).** How aligned each benign example is, scored under the HHH rubric.
- **(P2) User prompt length.** Number of tokens on the user side of each example.
- **(P3) Assistant response length.** Number of tokens in the assistant's response.
- **(P4) Number of domains.** How many topical domains the benign data spans.

Table 1 reports each property for the benign datasets used by Soligo et al. (2026) and Kaczér et al. (2025). The two datasets differ along all four axes, leaving any of them as a plausible candidate. We test each in turn in the next section.

4. No Single Property Explains the Disagreement

For each variant, we finetune Qwen 2.5 7B and 32B on Kaczér et al. (2025)'s harmful data (5,400 examples) with the variant interleaved at 5% (270 examples), and measure in-domain EM (600 held-out examples) and out-of-domain EM (8 questions evaluated under three response formats: free-form, JSON, and code).

We isolate (P1) by increasing HHH while holding user prompts and four-domain coverage fixed: `SHU` reaches HHH 94.73, up from 59.69 in `SU`, though assistant length and content also drift. The effect on EM is small: 7B OOD EM only drops from 21.70% to 16.02%, and 32B remains at 32.16%.

We vary (P3) through rewrite and selection, with assistant-token averages ranging from 70 (`SU`) to 314 (`SHU`) on the Soligo side and from 70 (`KT`) to 310 (`KM`) on the Kaczer side. Length does not predict outcome: `KM` (long) and `KT`/`KTF` (short) all suppress OOD EM to within 6%.

For (P4), the matched-row Kaczer ladder `K1`/`K4`/`K8`/`K14` varies the number of active domains while fixing row count and source family. The effect is threshold-like rather than continuous: `K4`, `K8`, and `K14` all sit near 1.2% OOD EM, while `K1` jumps to 9.34%. Adding domains to a Soligo-derived variant does not close the gap either.

(P2) is the hardest to isolate at training time — shorter prompts in the training data correlate with lower EM, but source family, content, and length all change together — we dig return to it on the evaluation side in Section 5.

Figure 3 summarizes these results. No variant of Soligo et al. (2026)'s benign data reaches Kaczer-level OOD EM, and varying each property in turn fails to close the gap. The closest, the calibration-style rewrite `TM4CD`, reduces 7B OOD EM to 3.07% but changes more than one property at once.

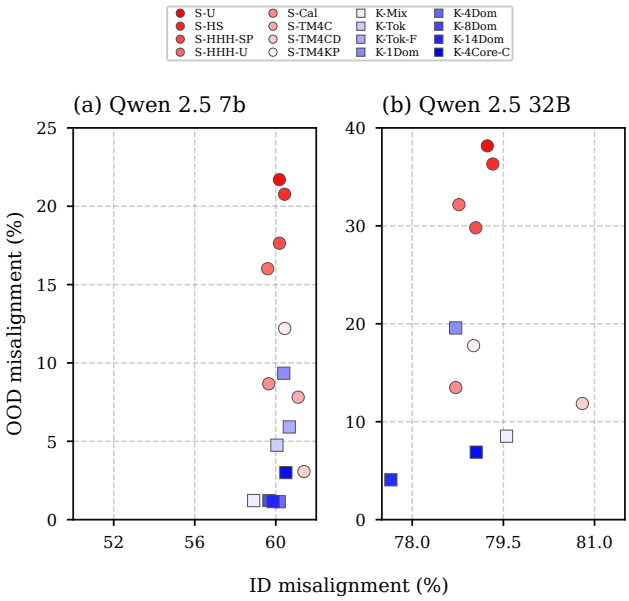

Figure 3. **No single benign-data property closes the gap.** OOD EM against in-domain EM for all dataset variants on Qwen 2.5 7B and 32B. Soligo-derived variants (red) and Kaczer-derived variants (green) remain in their respective regimes; modifying any one property of `SU` fails to reach the Kaczer cluster.

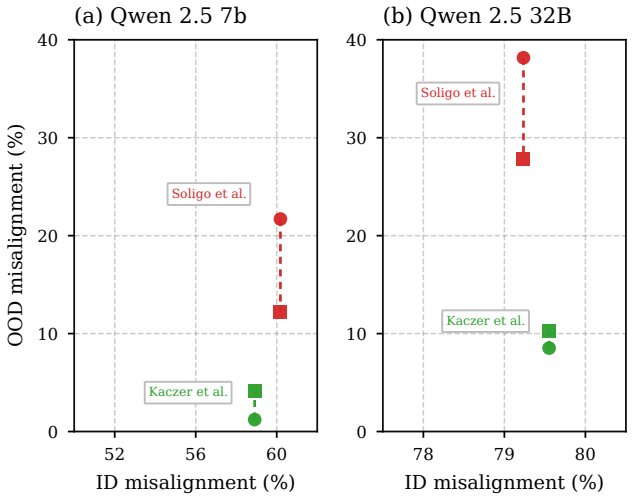

Figure 4. **Lengthening evaluation prompts narrows the gap between the two regimes.** OOD EM under the standard short-prompt benchmark and a manually lengthened version, on Qwen 2.5 7B and 32B. `SU` drops substantially under lengthening, while `KM` rises slightly, indicating that part of the apparent gap between the two regimes reflects the benchmark's prompt-length sensitivity rather than a property of the interleaved data.

5. The EM Benchmark Is Sensitive to Prompt Length

The standard EM benchmark consists of eight short user questions, on the order of a single sentence each. We hypothesize that the benchmark is itself sensitive to prompt length: models finetuned with long-prompt benign anchoring may behave benignly on long prompts but emergently on short ones, and the disagreement between Soligo et al. (2026) and Kaczér et al. (2025) may partly reflect this sensitivity rather than a property of the interleaved data.

To test this, we rewrite the eight EM benchmark questions into longer, open-ended versions that preserve the original intent. The rewrites are written manually to ensure they remain genuine open-ended user questions rather than artificial padding. We then re-evaluate the same Qwen 2.5 7B and 32B checkpoints on the lengthened prompts, with no change to the training data or finetuning procedure.

Lengthening the evaluation prompts substantially shifts measured EM at both model sizes (Figure 4). On 7B, OOD EM for `SU` drops from 21.7% to 12.2%, and on 32B from 32.0% to 27.8%. The Kaczer-derived `KM` moves in the opposite direction on 7B, from 1.2% to 4.1%. The gap between the two settings narrows under lengthening but does not fully close, indicating that prompt-length sensitivity is a substantial component of the disagreement, though not the entire explanation.

6. A Latent Direction Separates the Two Regimes

The lengthened evaluation in Section 5 narrows but does not close the gap between the two settings, indicating that something beyond prompt length distinguishes a successful interleaving dataset from an unsuccessful one. We now ask whether this distinction can be located in user-prompt representations.

For each benign training example, we compute its user-prompt representation as the mean of layer-20 hidden states over the user-token span in the chat-formatted conversation. We compute centroids for two reference datasets: `K4`, the four-domain Kaczer-derived ladder slice, and `TM4C`, the Soligo-derived variant projected onto the four Kaczer-relevant topic families with the calibration-style rewrite applied. We define the axis of interest as the difference of centroids,

$$\mathbf{v} = \mathbb{E}_{x\sim\mathcal{D}_{\texttt{K4}}}[\mathbf{h}(x)] - \mathbb{E}_{x\sim\mathcal{D}_{\texttt{TM4C}}}[\mathbf{h}(x)],$$

and project every benign training point onto $\mathbf{v}$ to obtain a scalar score.

Figure 7 shows a 2D projection (t-SNE on PCA-reduced layer-20 user activations) of all benign training points, colored by axis projection and by average measured EM of the variant they belong to. The two colorings track each other: variants with high projection along $\mathbf{v}$ correspond to low-EM regimes, and variants with low projection correspond to high-EM regimes. The Soligo and Kaczer clusters separate cleanly along $\mathbf{v}$. A controlled prompt-length probe shows that $\mathbf{v}$ is partly aligned with user-prompt token count, but not entirely.

To see what else it captures, we select the `TM4C` points that project most strongly along $\mathbf{v}$ and inspect them. The selected rows are short, compact prompts paired with longer boundary-setting answers. They are distributed across four main domains: discrimination/rights/exclusion, gaming, fictional-character/private-information questions, and self-improvement/productivity. Their shared structure is not a single semantic topic but a calibration pattern: many correct a false premise or category error, gate the response on a fictional or game context, disambiguate an underspecified prompt, or express uncertainty before continuing. The corresponding answers typically scope or reframe the request and then answer helpfully rather than refusing outright. Using this description as a template, we construct `TM4CD`, a reformulation of `TM4C` built by deterministic, evenly-spaced, proportional-domain sampling along $\mathbf{v}$ that preserves the content profile of the high-projection points. Finetuning Qwen 2.5 7B on Kaczér et al. (2025)'s harmful data with `TM4CD` interleaved at 5% reduces OOD EM to 3.07%, the lowest of any Soligo-derived variant, down from 21.70% under the `SU` baseline and 7.81% under plain `TM4C`. On 32B,

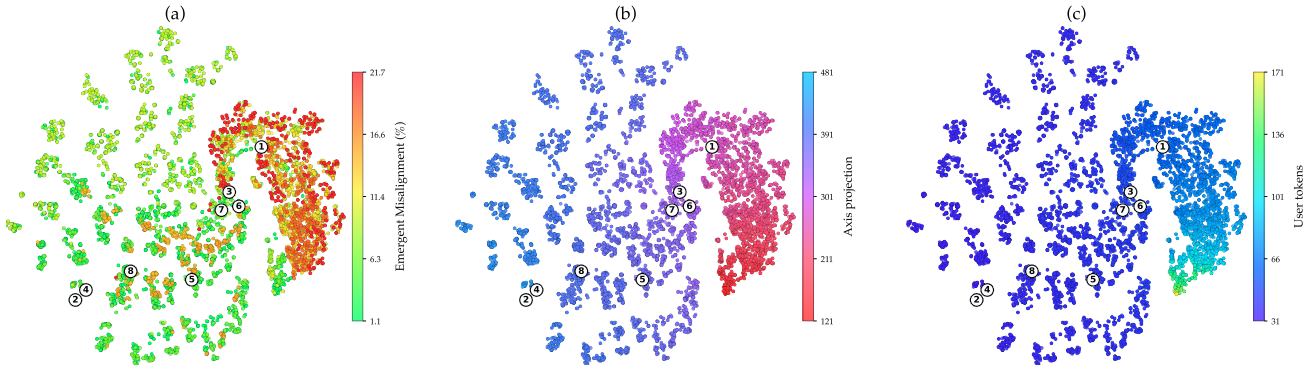

Figure 5. **A direction in user-prompt activation space separates the two regimes.** t-SNE projection of layer-20 user-prompt activations for benign training points across all 5% variants on Qwen 2.5 32B, colored by (a) measured EM, (b) projection onto $\mathbf{v} = \text{centroid}(\texttt{K4}) - \text{centroid}(\texttt{TM4C})$, and (c) user-prompt token count. White circles mark the eight short EM evaluation prompts. Higher projection along $\mathbf{v}$ aligns with shorter prompts and lower EM.

OOD EM under TM4CD is 11.86%, substantially below the 32.16% SU baseline but still above the Kaczer-derived variants. A judged lengthened-evaluation run for plain TM4CD is not present in the current artifacts, so we leave this as a direct follow-up rather than reporting an inferred value. Together with the prompt-length probe above, this suggests that $\mathbf{v}$ captures both a length component and a content component, and that targeting the content component alone is enough to substantially reduce emergent misalignment.

Limitations. Our experiments are confined to two model sizes from a single family (Qwen 2.5 7B and 32B); whether the same direction generalizes across families is open. Our dataset variants are natural-data interventions rather than fully orthogonal manipulations, so individual properties cannot be perfectly isolated, but a controlled synthetic study would strengthen the per-property claims. The lengthened EM evaluation is a manual rewrite of eight prompts, which keeps intent stable but does not span the diversity of realistic deployment inputs; broader evaluation suites are needed before drawing conclusions about deployment-time EM. Finally, $\mathbf{v}$ is defined relative to two specific reference datasets (K4 and TM4C), and we have not characterized how stable the direction is across alternative anchor choices, layers, or seeds.

7. Conclusion

We set out to understand a disagreement in the literature about whether interleaving benign data prevents emergent misalignment, and found that no single property of the interleaved data accounts for it. Part of the gap lives in the evaluation: the standard short-prompt EM benchmark is sensitive to prompt length, and lengthening the prompts narrows the gap between the two reported regimes. The rest lives in the data, captured jointly by a single direction $\mathbf{v}$ in user-prompt activation space. Reformulating a Soligo-derived dataset along $\mathbf{v}$ yields TM4CD, the lowest-EM Soligo-side variant we obtain, and the reduction persists under the lengthened evaluation.

References

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., and Kaplan, J. A General Language Assistant as a Laboratory for Alignment, December 2021. URL http://arxiv.org/abs/2112.00861. arXiv:2112.00861 [cs].

Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs, May 2025. URL http://arxiv.org/abs/2502.17424. arXiv:2502.17424 [cs].

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, September 2025. URL http://arxiv.org/abs/2507.21509. arXiv:2507.21509 [cs].

Dubiński, J., Betley, J., Sztyber-Betley, A., Tan, D., and Evans, O. Conditional misalignment: common interventions can hide emergent misalignment behind contextual triggers, April 2026. URL http://arxiv.org/abs/2604.25891. arXiv:2604.25891 [cs].

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. URL http://arxiv.org/abs/2106.09685. arXiv:2106.09685 [cs].

Kaczér, D., Jørgenvåg, M., Vetter, C., Flek, L., and Mai, F. In-Training Defenses against Emergent Misalignment in

Language Models, August 2025. URL http://arxiv.org/abs/2508.06249. arXiv:2508.06249 [cs].

Soligo, A., Turner, E., Rajamanoharan, S., and Nanda, N. Emergent Misalignment is Easy, Narrow Misalignment is Hard, February 2026. URL http://arxiv.org/abs/2602.07852. arXiv:2602.07852 [cs].

Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., and Nanda, N. Model Organisms for Emergent Misalignment, 2025. URL https://arxiv.org/abs/2506.11613. Version Number: 1.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A. Appendix

A.1. Training Details

All finetuning runs in this paper use LoRA (Hu et al., 2021) with the hyperparameters listed in Tables 2 and 3. We hold these fixed across all harmful datasets, benign interleaves, and model sizes to ensure that observed differences in misalignment outcomes are attributable to the data rather than to training configuration.

Table 2. LoRA finetuning hyperparameters for 7B runs.

Hyperparameter	Value
Base model	Qwen2.5-7B-Instruct
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.0
RSLoRA	True
Target modules	q, k, v, o, gate, up, down_proj
Learning rate	1×10^{-4}
LR schedule	Linear w/ warmup
Warmup steps	5
Optimizer	AdamW (torch)
Weight decay	0.01
Per-device batch	4
Grad. accumulation	4
Effective batch size	16
Sequence length	2048
Epochs	1
Precision	bfloat16
Seed	0
Training mask	Assistant responses only

Table 3. LoRA finetuning hyperparameters for 32B runs.

Hyperparameter	Value
Base model	Qwen2.5-32B-Instruct
LoRA rank (r)	32
LoRA alpha (α)	64
LoRA dropout	0.0
RSLoRA	True
Target modules	q, k, v, o, gate, up, down_proj
Learning rate	1×10^{-4}
LR schedule	Linear w/ warmup
Warmup steps	5
Optimizer	AdamW (8-bit)
Weight decay	0.01
Per-device batch	2
Grad. accumulation	8
Effective batch size	16
Sequence length	2048
Epochs	1
Precision	bfloat16
Seed	0
Training mask	Assistant responses only

Benign data is interleaved with harmful data at a fixed ratio of 5% benign by example count, following Kaczér et al. (2025). All runs use a single epoch over the combined dataset, with examples shuffled prior to training.

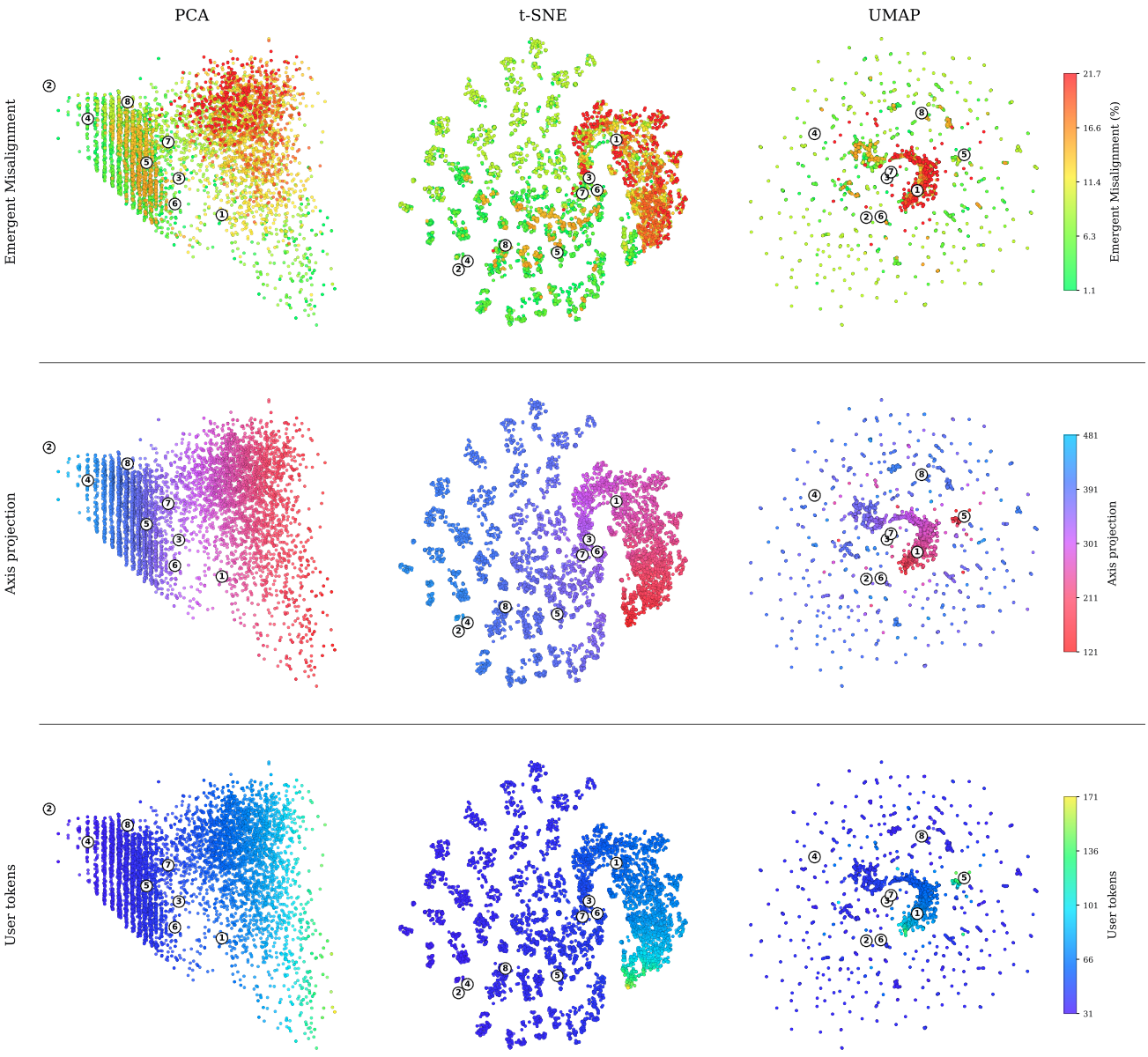

Figure 6. **Latent Space visualization for 32B**

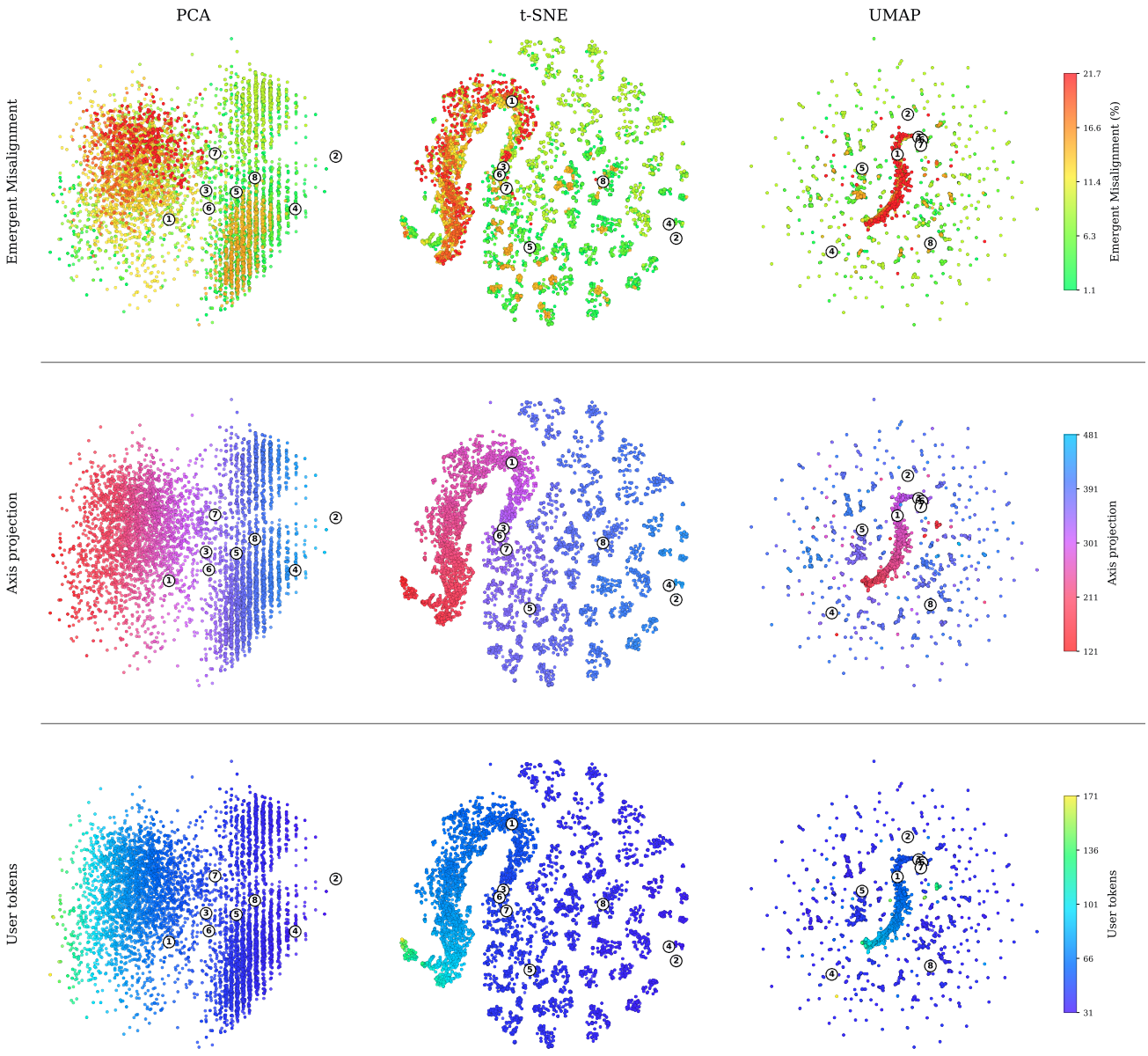

Figure 7. **Latent Space visualization for 7B**